

## Assumptions for Inference

## And the Conditions That Support or Override Them

### Proportions ( $z$ )

- **One sample**

1. Individuals are independent.
2. Sample is sufficiently large.

1. SRS and  $n < 10\%$  of the population.
2. Successes and failures each  $\geq 10$ .

- **Two groups**

1. Groups are independent.
2. Data in each group are independent.
3. Both groups are sufficiently large.

1. (Think about how the data were collected.)
2. Both are SRSs and  $n < 10\%$  of populations OR random allocation.
3. Successes and failures each  $\geq 10$  for both groups.

### Means ( $t$ )

- **One Sample** ( $df = n - 1$ )

1. Individuals are independent.
2. Population has a Normal model.

1. SRS and  $n < 10\%$  of the population.
2. Histogram is unimodal and symmetric.\*

- **Matched pairs** ( $df = n - 1$ )

1. Data are matched.
2. Individuals are independent.
3. Population of differences is Normal.

1. (Think about the design.)
2. SRS and  $n < 10\%$  OR random allocation.
3. Histogram of differences is unimodal and symmetric.\*

- **Two independent groups** ( $df$  from technology)

1. Groups are independent.
2. Data in each group are independent.
3. Both populations are Normal.

1. (Think about the design.)
2. SRSs and  $n < 10\%$  OR random allocation.
3. Both histograms are unimodal and symmetric.\*

### Distributions/Association ( $\chi^2$ )

- **Goodness of fit** ( $df = \#$  of cells  $- 1$ ; one variable, one sample compared with population model)

1. Data are counts.
2. Data in sample are independent.
3. Sample is sufficiently large.

1. (Are they?)
2. SRS and  $n < 10\%$  of the population.
3. All expected counts  $\geq 5$ .

- **Homogeneity** [ $df = (r - 1)(c - 1)$ ; many groups compared on one variable]

1. Data are counts.
2. Data in groups are independent.
3. Groups are sufficiently large.

1. (Are they?)
2. SRSs and  $n < 10\%$  OR random allocation.
3. All expected counts  $\geq 5$ .

- **Independence** [ $df = (r - 1)(c - 1)$ ; sample from one population classified on two variables]

1. Data are counts.
2. Data are independent.
3. Sample is sufficiently large.

1. (Are they?)
2. SRSs and  $n < 10\%$  of the population.
3. All expected counts  $\geq 5$ .

### Regression ( $t$ , $df = n - 2$ )

- **Association** between two quantitative variables ( $\beta = 0$ ?)

1. Form of relationship is linear.
2. Errors are independent.
3. Variability of errors is constant.
4. Errors have a Normal model.

1. Scatterplot looks approximately linear.
2. No apparent pattern in residuals plot.
3. Residuals plot has consistent spread.
4. Histogram of residuals is approximately unimodal and symmetric, or normal probability plot reasonably straight.\*

(\*less critical as  $n$  increases)

## Quick Guide to Inference

Think		Show					Tell?
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
<b>Proportions</b>	One sample	1-Proportion z-Interval	$z$	$p$	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion z-Test				$\sqrt{\frac{p_0q_0}{n}}$	20, 21
	Two independent groups	2-Proportion z-Interval	$z$	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion z-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
<b>Means</b>	One sample	$t$ -Interval $t$ -Test	$t$ $df = n - 1$	$\mu$	$\bar{y}$	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample $t$ -Test 2-Sample $t$ -Interval	$t$ $df$ from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	Matched pairs	Paired $t$ -Test Paired $t$ -Interval	$t$ $df = n - 1$	$\mu_d$	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	25
<b>Distributions</b> (one categorical variable)	One sample	Goodness-of-Fit	$\chi^2$ $df = \text{cells} - 1$	$\sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$			26
	Many independent groups	Homogeneity $\chi^2$ Test	$\chi^2$ $df = (r - 1)(c - 1)$				
<b>Independence</b> (two categorical variables)	One sample	Independence $\chi^2$ Test					
<b>Association</b> (two quantitative variables)	One sample	Linear Regression $t$ -Test or Confidence Interval for $\beta$	$t$ $df = n - 2$	$\beta_1$	$b_1$	$\frac{s_e}{s_x \sqrt{n - 1}}$ (compute with technology)	27
		*Confidence Interval for $\mu_v$		$\mu_v$	$\hat{y}_v$	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n}}$	
		*Prediction Interval for $y_v$		$y_v$	$\hat{y}_v$	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter

# Stats

**Modeling the World**

THIRD EDITION



EDITION

3

# Stats

## Modeling the World

**David E. Bock**

Ithaca High School  
Cornell University

**Paul F. Velleman**

Cornell University

**Richard D. De Veaux**

Williams College

**Addison-Wesley**

Boston San Francisco New York

London Toronto Sydney Tokyo Singapore Madrid

Mexico City Munich Paris Cape Town Hong Kong Montreal

<i>Editor in Chief</i>	Deirdre Lynch
<i>Acquisitions Editor</i>	Christopher Cummings
<i>Senior Editor, AP and Electives</i>	Andrea Sheehan
<i>Assistant Editor</i>	Christina Lepre
<i>Editorial Assistant</i>	Dana Jones
<i>Senior Project Editor</i>	Chere Bemelmans
<i>Senior Managing Editor</i>	Karen Wernholm
<i>Senior Production Supervisor</i>	Sheila Spinney
<i>Cover Design</i>	Barbara T. Atkinson
<i>Digital Assets Manager</i>	Marianne Groth
<i>Media Producer</i>	Christine Stavrou
<i>Software Development</i>	Edward Chappell (MathXL) and Marty Wright (TestGen)
<i>Marketing Manager</i>	Alex Gay
<i>Marketing Coordinator</i>	Kathleen DeChavez
<i>Senior Author Support/Technology Specialist</i>	Joe Vetere
<i>Senior Prepress Supervisor</i>	Caroline Fell
<i>Senior Manufacturing Buyer</i>	Carol Melville
<i>Senior Media Buyer</i>	Ginny Michaud
<i>Production Coordination, Composition, and Illustrations</i>	Pre-Press PMG
<i>Interior Design</i>	The Davis Group, Inc.
<i>Cover Photo</i>	Pete McArthur

#### Library of Congress Cataloging-in-Publication Data

Bock, David E.

Stats : modeling the world / David E. Bock, Paul F. Velleman, Richard D. De Veaux.— 3rd ed.

p. cm.

Includes index.

ISBN 13: 978-0-13-135958-1

ISBN 10: 0-13-135958-4

1. Graphic calculators—Textbooks. I. Velleman, Paul F., 1949- II. De Veaux, Richard D. III. Title.

QA276.12.B628 2010

519.5—dc22

2008029019

For permission to use copyrighted material, grateful acknowledgement has been made to the copyright holders listed in Appendix D, which is hereby made part of this copyright page.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial caps or all caps. TI-Nspire and the TI-Nspire logo are trademarks of Texas Instruments, Inc.

Copyright © 2010, 2007, 2004 Pearson Education, Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. For information on obtaining permission for use of material in this work, please submit a written request to Pearson Education, Inc., Rights and Contracts Department, 501 Boylston Street, Boston, MA 02116, fax your request to 617-848-7047, or e-mail at <http://www.pearsoned.com/legal/permissions.htm>.

1 2 3 4 5 6 7 8 9 10—CRK—12 11 10 09

**Addison-Wesley**  
is an imprint of



[www.PearsonSchool.com/Advanced](http://www.PearsonSchool.com/Advanced)

ISBN 13: 978-0-13-135958-1

ISBN 10: 0-13-135958-4

*To Greg and Becca, great fun as kids and great friends as adults,  
and especially to my wife and best friend, Joanna, for her  
understanding, encouragement, and love*

*—Dave*

*To my sons, David and Zev, from whom I've learned so much,  
and to my wife, Sue, for taking a chance on me*

*—Paul*

*To Sylvia, who has helped me in more ways than she'll ever know,  
and to Nicholas, Scyrine, Frederick, and Alexandra,  
who make me so proud in everything that they are and do*

*—Dick*

# Meet the Authors



**David E. Bock** taught mathematics at Ithaca High School for 35 years. He has taught Statistics at Ithaca High School, Tompkins-Cortland Community College, Ithaca College, and Cornell University. Dave has won numerous teaching awards, including the MAA's Edyth May Sliffe Award for Distinguished High School Mathematics Teaching (twice), Cornell University's Outstanding Educator Award (three times), and has been a finalist for New York State Teacher of the Year.

Dave holds degrees from the University at Albany in Mathematics (B.A.) and Statistics/Education (M.S.). Dave has been a reader and table leader for the AP Statistics exam, serves as a Statistics consultant to the College Board, and leads workshops and institutes for AP Statistics teachers. He has recently served as K–12 Education and Outreach Coordinator and a senior lecturer for the Mathematics Department at Cornell University. His understanding of how students learn informs much of this book's approach.

Dave relaxes by biking and hiking. He and his wife have enjoyed many days camping across Canada and through the Rockies. They have a son, a daughter, and three grandchildren.



**Paul F. Velleman** has an international reputation for innovative Statistics education. He is the author and designer of the multimedia statistics CD-ROM *ActivStats*, for which he was awarded the EDUCOM Medal for innovative uses of computers in teaching statistics, and the ICTCM Award for Innovation in Using Technology in College Mathematics. He also developed the award-winning statistics program, Data Desk, and the Internet site Data And Story Library (DASL) (<http://dasl.datadesk.com>), which provides data sets for teaching Statistics. Paul's understanding of using and teaching with technology informs much of this book's approach.

Paul has taught Statistics at Cornell University since 1975. He holds an A.B. from Dartmouth College in Mathematics and Social Science, and M.S. and Ph.D. degrees in Statistics from Princeton University, where he studied with John Tukey. His research often deals with statistical graphics and data analysis methods. Paul co-authored (with David Hoaglin) *ABCs of Exploratory Data Analysis*. Paul is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science.

Out of class, Paul sings baritone in a barbershop quartet. He is the father of two boys.



**Richard D. De Veaux** is an internationally known educator and consultant. He has taught at the Wharton School and the Princeton University School of Engineering, where he won a "Lifetime Award for Dedication and Excellence in Teaching." Since 1994, he has been Professor of Statistics at Williams College. Dick has won both the Wilcoxon and Shewell awards from the American Society for Quality. He is a fellow of the American Statistical Association. Dick is also well known in industry, where for the past 20 years he has consulted for such companies as Hewlett-Packard, Alcoa, DuPont, Pillsbury, General Electric, and Chemical Bank. He has also sometimes been called the "Official Statistician for the Grateful Dead." His real-world experiences and anecdotes illustrate many of this book's chapters.

Dick holds degrees from Princeton University in Civil Engineering (B.S.E.) and Mathematics (A.B.) and from Stanford University in Dance Education (M.A.) and Statistics (Ph.D.), where he studied with Persi Diaconis. His research focuses on the analysis of large data sets and data mining in science and industry.

In his spare time he is an avid cyclist and swimmer. He also is the founder and bass for the "Diminished Faculty," an a cappella Doo-Wop quartet at Williams College. Dick is the father of four children.

# Contents

Preface ix



## Exploring and Understanding Data 1

- CHAPTER 1 Stats Start Here 2
- CHAPTER 2 Data 7
- CHAPTER 3 Displaying and Describing Categorical Data 20
- CHAPTER 4 Displaying and Summarizing Quantitative Data 44
- CHAPTER 5 Understanding and Comparing Distributions 80
- CHAPTER 6 The Standard Deviation as a Ruler and the Normal Model 104
- Review of Part I Exploring and Understanding Data 135



## Exploring Relationships Between Variables 145

- CHAPTER 7 Scatterplots, Association, and Correlation 146
- CHAPTER 8 Linear Regression 171
- CHAPTER 9 Regression Wisdom 201
- CHAPTER 10 Re-expressing Data: Get It Straight! 222
- Review of Part II Exploring Relationships Between Variables 244



## Gathering Data 253

- CHAPTER 11 Understanding Randomness 255
- CHAPTER 12 Sample Surveys 268
- CHAPTER 13 Experiments and Observational Studies 292
- Review of Part III Gathering Data 317



## Randomness and Probability 323

- CHAPTER 14 From Randomness to Probability 324
- CHAPTER 15 Probability Rules! 342
- CHAPTER 16 Random Variables 366
- CHAPTER 17 Probability Models 388
- Review of Part IV Randomness and Probability 405





## From the Data at Hand to the World at Large 411

**CHAPTER 18** Sampling Distribution Models 412

**CHAPTER 19** Confidence Intervals for Proportions 439

**CHAPTER 20** Testing Hypotheses About Proportions 459

**CHAPTER 21** More About Tests and Intervals 480

**CHAPTER 22** Comparing Two Proportions 504

**Review of Part V** From the Data at Hand to the World at Large 523



## Learning About the World 529

**CHAPTER 23** Inferences About Means 530

**CHAPTER 24** Comparing Means 560

**CHAPTER 25** Paired Samples and Blocks 587

**Review of Part VI** Learning About the World 609



## Inference When Variables Are Related 617

**CHAPTER 26** Comparing Counts 618

**CHAPTER 27** Inferences for Regression 649

**Review of Part VII** Inference When Variables Are Related 683

**CHAPTER 28** \*Analysis of Variance—on the DVD

**CHAPTER 29** \*Multiple Regression—on the DVD

## Appendixes

**A** Selected Formulas A-1

**B** Guide to Statistical Software A-3

**C** Answers A-25

**D** Photo Acknowledgments A-59

**E** Index A-61

**F** TI Tips A-71

**G** Tables A-73

---

\*Indicates an optional chapter.

# Preface

## About the Book

**W**e've been thrilled with the feedback we've received from teachers and students using *Stats: Modeling the World*, Second Edition. If there is a single hallmark of this book it is that students actually read it. We have reports from every level—from high school to graduate school—that students find our books easy and even enjoyable to read. We strive for a conversational, approachable style, and introduce anecdotes to maintain students' interest. And it works. Teachers report their amazement that students are voluntarily reading ahead of their assignments. Students write to tell us (to their amazement) that they actually enjoyed the book.

*Stats: Modeling the World*, Third Edition is written from the ground up with the understanding that Statistics is practiced with technology. This insight informs everything from our choice of forms for equations (favoring intuitive forms over calculation forms) to our extensive use of real data. Most important, it allows us to focus on teaching Statistical Thinking rather than calculation. The questions that motivate each of our hundreds of examples are not “how do you find the answer?” but “how do you think about the answer?”

## Our Goal: Read This Book!

The best text in the world is of little value if students don't read it. Here are some of the ways we have made *Stats: Modeling the World*, Third Edition even more approachable:

- **Readability.** You'll see immediately that this book doesn't read like other Statistics texts. The style, both colloquial (with occasional humor) and informative, engages students to actually read the book to see what it says.
- **Informality.** Our informal diction doesn't mean that the subject matter is covered lightly or informally. We have tried to be precise and, wherever possible, to offer deeper explanations and justifications than those found in most introductory texts.
- **Focused lessons.** The chapters are shorter than in most other texts, to make it easier to focus on one topic at a time.
- **Consistency.** We've worked hard to avoid the “do what we say, not what we do” trap. From the very start we teach the importance of plotting data and checking assumptions and conditions, and we have been careful to model that behavior right through the rest of the book.
- **The need to read.** Students who plan just to skim the book may find our presentation a bit frustrating. The important concepts, definitions, and sample solutions don't sit in little boxes. This is a book that needs to be read, so we've tried to make the reading experience enjoyable.

## New to the Third Edition

The third edition of *Stats: Modeling the World* continues and extends the successful innovations pioneered in our books, teaching Statistics and statistical thinking as it is practiced today. We've rewritten sections throughout the book to make them clearer and more interesting. We've introduced new up-to-the-minute motivating examples throughout. And, we've added a number of new features, each with the goal of making it even easier for students to put the concepts of Statistics together into a coherent whole.

### FOR EXAMPLE

- ▶ **For Example.** In every chapter, you'll find approximately 4 new worked examples that illustrate how to apply new concepts and methods—**more than 100 new illustrative examples**. But these aren't isolated examples. We carry a discussion through the chapter with each *For Example*, picking up the story and moving it forward as students learn to apply each new concept.

### STEP-BY-STEP EXAMPLE

- ▶ **Step-by-Step Worked Examples.** We've brought our innovative *Think/Show/Tell Step-by-Step* examples up-to-date with new examples and data.

### A S

- ▶ **ActivStats Pointers.** In the third edition, the *ActivStats* pointers have been revised for clarity and now indicate exactly what they are pointing to—activity, video, simulation, or animation—paralleling the book's discussions to enhance learning.

### TI-Nspire

- ▶ **TI-Nspire Activities.** We've created many demonstrations and investigations for TI-Nspire handhelds to enhance each chapter. They're on the DVD and at the book's Web site.

- ▶ **Exercises.** We've added **hundreds of new exercises**, including more single-concept exercises at the beginning of each set so students can be sure they have a clear understanding of each important topic before they're asked to tie them all together in more comprehensive exercises. Continuing exercises have been **updated with the most recent data**. Whenever possible, the data are on the DVD and the book's Web site so students can explore them further.

- ▶ **Data Sources.** Most of the data used in examples and exercises are from recent news stories, research articles, and other real-world sources. We've listed more of those sources in this edition.

- ▶ **Chapters 4 and 5** have been entirely rewritten and reorganized. We think you'll agree with our reviewers that the new organization—discussing displays and summaries for quantitative data in Chapter 4 and then expanding on those ideas to discuss comparisons across groups, outliers, and other more sophisticated topics in Chapter 5—provides a more exciting and interesting way to approach these fundamental topics.

- ▶ **Simulation.** We've improved the discussion of simulation in Chapter 11 so it could relate more easily to discussions of experimental design and probability. The simulations included in the *ActivStats* multimedia software on the book's DVD carry those ideas forward in a student-friendly fashion.

- ▶ **Teacher's Podcasts** (10 points in 10 minutes). Created and presented by the authors, these podcasts focus on key points in each chapter to help you with class preparation. These podcasts are available on the Instructor's Resource CD.

- ▶ **Video Lectures on DVD** featuring the textbook authors will help students review the high points of each chapter. Video presenters also work through examples from the text. The presentations feature the same student-friendly style and emphasis on critical thinking as the text.

## Continuing Features



▶ *Think, Show, Tell.* The worked examples repeat the mantra of *Think, Show,* and *Tell* in every chapter. They emphasize the importance of thinking about a Statistics question (What do we know? What do we hope to learn? Are the assumptions and conditions satisfied?) and reporting our findings (the *Tell* step). The *Show* step contains the mechanics of calculating results and conveys our belief that it is only one part of the process. This rubric is highlighted in the *Step-by-Step* examples that guide the students through the process of analyzing the problem with the general explanation on the left and the worked-out problem on the right. The result is a better understanding of the concept, not just number crunching.



### JUST CHECKING

▶ *Just Checking.* Within each chapter, we ask students to pause and think about what they've just read. These questions are designed to be a quick check that they understand the material. Answers are at the end of the exercise sets in each chapter so students can easily check themselves.



▶ *TI Tips.* We emphasize sound understanding of formulas and methods, but want students to use technology for actual calculations. Easy-to-read “TI Tips” in the chapters show students how to use TI-83/84 Plus statistics functions. (Help using a TI-89 or TI-Nspire appears in Appendix B.) We do remind students that calculators are just for “Show”—they cannot Think about what to do nor Tell what it all means.



▶ *Math Boxes.* In many chapters we present the mathematical underpinnings of the statistical methods and concepts. By setting these proofs, derivations, and justifications apart from the narrative, we allow the student to continue to follow the logical development of the topic at hand, yet also refer to the underlying mathematics for greater depth.



▶ *What Can Go Wrong?* Each chapter still contains our innovative *What Can Go Wrong?* sections that highlight the most common errors people make and the misconceptions they have about Statistics. Our goals are to help students avoid these pitfalls, and to arm them with the tools to detect statistical errors and to debunk misuses of statistics, whether intentional or not. In this spirit, some of our exercises probe the understanding of such failures.



▶ *What Have We Learned?* These chapter-ending summaries are great study guides providing complete overviews that highlight the new concepts, define the new terms, and list the skills that the student should have acquired in the chapter.

▶ *Exercises.* Throughout, we've maintained the pairing of examples so that each odd-numbered exercise (with an answer in the back of the book) is followed by an even-numbered exercise on the same concept. Exercises are still ordered by level of difficulty.



▶ *Reality Check.* We regularly remind students that Statistics is about understanding the world with data. Results that make no sense are probably wrong, no matter how carefully we think we did the calculations. Mistakes are often easy to spot with a little thought, so we ask students to stop for a reality check before interpreting their result.

### NOTATION ALERT:

▶ *Notation Alert.* Throughout this book we emphasize the importance of clear communication, and proper notation is part of the vocabulary of Statistics. We've found that it helps students when we call attention to the letters and symbols statisticians use to mean very specific things.




---

**ON THE COMPUTER**


---

- ▶ **Connections.** Each chapter has a *Connections* section to link key terms and concepts with previous discussions and to point out continuing themes, helping students fit newly learned concepts into a growing understanding of Statistics.
- ▶ **On the Computer.** In the real world, Statistics is practiced with computers. We prefer not to choose a particular Statistics program. Instead, at the end of each chapter, we summarize what students can find in the most common packages, often with an annotated example. Computer output appearing in the book and in exercises is often generic, resembling all of the common packages to some degree.

## Coverage

Textbooks are often defined more by what they choose not to cover than by what they do cover. We've been guided in the choice and order of topics by several fundamental principles. First, we have tried to ensure that each new topic fits into the growing structure of understanding that we hope students will build. Several topic orders can support this goal. We explain our reasons for the topic order of the chapters in the ancillary Printed Test Bank and Resource Guide.

**GAISE Guidelines.** We have worked to provide materials to help each class, in its own way, follow the guidelines of the GAISE (Guidelines for Assessment and Instruction in Statistics Education) project sponsored by the American Statistical Association. That report urges that Statistics education should

1. emphasize Statistical literacy and develop Statistical thinking,
2. use real data,
3. stress conceptual understanding rather than mere knowledge of procedures,
4. foster active learning,
5. use technology for developing concepts and analyzing data, and
6. make assessment a part of the learning process.

We also have been guided by the syllabus of the AP\* Statistics course. We agree with the wisdom of those who designed that course in their selection of topics and their emphasis on Statistics as a practical discipline. *Stats: Modeling the World* provides complete discussions of all AP\* topics and teaches students communication skills that lead to success on the AP\* examination. A correlation of the text to the AP\* Statistics course standards is available in the Printed Test Bank and Resource Guide, on the Instructor's Resource CD, and at [www.phschool.com/advanced/correlations/statistics.html](http://www.phschool.com/advanced/correlations/statistics.html).

## Mathematics

Mathematics traditionally appears in Statistics texts in several roles:

1. It can provide a concise, clear statement of important concepts.
2. It can describe calculations to be performed with data.
3. It can embody proofs of fundamental results.

Of these, we emphasize the first. Mathematics can make discussions of Statistics concepts, probability, and inference clear and concise. We have tried to be sensitive to those who are discouraged by equations by also providing verbal descriptions and numerical examples.

This book is not concerned with proving theorems about Statistics. Some of these theorems are quite interesting, and many are important. Often, though, their proofs are not enlightening to introductory Statistics students, and can distract the audience from the concepts we want them to understand. However, we have not shied

away from the mathematics where we believed that it helped clarify without intimidating. You will find some important proofs, derivations, and justifications in Math Boxes that accompany the development of many topics.

Nor do we concentrate on calculations. Although statistics calculations are generally straightforward, they are also usually tedious. And, more to the point, they are often unnecessary. Today, virtually all statistics are calculated with technology, so there is little need for students to work by hand. The equations we use have been selected for their focus on understanding concepts and methods.

## Technology and Data

To experience the real world of Statistics, it's best to explore real data sets using modern technology.

- ▶ **Technology.** We assume that you are using some form of technology in your Statistics course. That could be a calculator, a spreadsheet, or a statistics package. Rather than adopt any particular software, we discuss generic computer output. "TI-Tips"—included in most chapters—show students how to use statistics features of the TI-83/84 Plus series. The Companion DVD, included in the Teacher's Edition, may be purchased for students and includes *ActivStats* and the software package Data Desk. Also, in Appendix B, we offer general guidance (by chapter) to help students get started on five common software platforms (Excel, MINITAB, Data Desk, JMP, and SPSS), a TI-89 calculator, and a TI-Nspire.
- ▶ **Data.** Because we use technology for computing, we don't limit ourselves to small, artificial data sets. In addition to including some small data sets, we have built examples and exercises on real data with a moderate number of cases—usually more than you would want to enter by hand into a program or calculator. These data are included on the DVD as well as on the book's Web site, [www.aw.com/bock](http://www.aw.com/bock).

### ON THE DVD

The DVD holds a number of supporting materials, including *ActivStats*, the *Data Desk* statistics package, an Excel add-in (DDXL), all large data sets from the text formatted for the most popular technologies, and two additional chapters.

***ActivStats (for Data Desk).*** The award-winning *ActivStats* multimedia program supports learning chapter by chapter. It complements the book with videos of real-world stories, worked examples, animated expositions of each of the major Statistics topics, and tools for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes

- improved navigation and a cleaner design that makes it easier to find and use tools such as the Index and Glossary
- more than **1000 homework exercises**, including many new exercises, plus answers to the "odd numbered" exercises. Many are from the text, providing the data already set up for calculations, and some are unique to *ActivStats*. Many exercises link to data files for each statistics package.
- **17 short video clips**, many new and updated
- **70 animated activities**
- **117 teaching applets**
- more than **300 data sets**

## Supplements

### STUDENT SUPPLEMENTS

The following supplements are available for purchase:

**Graphing Calculator Manual**, by Patricia Humphrey (Georgia Southern University) and John Diehl (Hinsdale Central High School), is organized to follow the sequence of topics in the text, and is an easy-to-follow, step-by-step guide on how to use the TI-83/84 Plus, TI-89, and TI-Nspire™ graphing calculators. It provides worked-out examples to help students fully understand and use the graphing calculator. (ISBN-13: 978-0-321-57094-9; ISBN-10: 0-321-57094-4)

**Pearson Education AP\* Test Prep Series: Statistics** by Anne Carroll, Ruth Carver, Susan Peters, and Janice Ricks, is written specifically to complement *Stats: Modeling the World, Third Edition, AP\* Edition*, and to help students prepare for the AP\* Statistics exam. Students can review topics that are discussed in *Stats: Modeling the World, Third Edition AP\* Edition*, and are likely to appear on the Advanced Placement Exam. The guide also contains test-taking strategies as well as practice tests. (ISBN 13: 978-0-13-135964-2; ISBN-10: 0-13-135964-9)

**Statistics Study Card** is a resource for students containing important formulas, definitions, and tables that correspond precisely to the De Veaux/Velleman/Bock Statistics series. This card can work as a reference for completing homework assignments or as an aid in studying. (ISBN-13: 978-0-321-46370-8; ISBN-10: 0-321-46370-6)

**Graphing Calculator Tutorial for Statistics** will guide students through the keystrokes needed to most efficiently use their graphing calculator. Although based on the TI-84 Plus Silver Edition, operating system 2.30, the keystrokes for this calculator are identical to those on the TI-84 Plus, and very similar to the TI-83 and TI-83 Plus. This tutorial should be helpful to students using any of these calculators, though there may be differences in some lessons. The tutorial is organized by topic. (ISBN-13: 978-0-321-41382-6; ISBN-10: 0-321-41382-2)

### TEACHER SUPPLEMENTS

Most of the teacher supplements and resources for this book are available electronically. On adoption or to preview, please go to [PearsonSchool.com/Advanced](http://PearsonSchool.com/Advanced) and click “Online Teacher Supplements.” You will be required to complete a one-time registration subject to verification before being emailed access information to download materials.

The following supplements are available to qualified adopters:

**Teacher’s Edition** contains answers to all exercises. Packaged with the Teacher’s Edition is the Companion DVD and the Instructor’s Resource CD. The Instructor’s Resource CD includes the Teachers’ Solutions Manual, Test Bank and Resource Guide, Audio Podcasts, PowerPoint slides, and Graphing Calculator Manual. (ISBN-13: 978-0-13-135959-8; ISBN-10: 0-13-135959-2)

**Printed Test Bank and Resource Guide**, by William Craine, contains chapter-by-chapter comments on the major concepts, tips on presenting topics (and what to avoid), teaching examples, suggested assignments, Web links and lists of other resources, as well as chapter quizzes, unit tests, investigative tasks, TI-Nspire activities, and suggestions for projects. An indispensable guide to help teachers prepare for class, the previous editions were soundly praised by new teachers of Statistics and seasoned veterans alike. The Printed Test Bank and Resource Guide is on the Instructor’s Resource CD and available for download. (ISBN-13: 978-0-13-135960-4; ISBN-10: 0-13-135960-6)

**Teacher’s Solutions Manual**, by William Craine, contains detailed solutions to all of the exercises. (ISBN-13: 978-0-13-136009-9; ISBN-10: 0-13-136009-4)

**TestGen® CD** enables teachers to build, edit, print, and administer tests using a computerized bank of questions developed to cover all the objectives of the text. TestGen is algorithmically based, allowing teachers to create multiple but equivalent versions of the same question or test with the click of a button. Teachers can also modify test bank questions or add new questions. Tests can be printed or administered online. (ISBN-13: 978-0-13-135961-1; ISBN-10: 0-13-135961-4)

**PowerPoint Lecture Slides** provide an outline to use in a lecture setting, presenting definitions, key concepts, and figures from the text. These slides are available on the Instructor’s Resource CD and available for download. (ISBN-13: 978-0-321-57101-4; ISBN 10: 0-321-57101-0)

## Technology Resources

**Instructor's Resource CD**, packaged with every new Teacher's Edition, includes the Teacher's Solutions Manual, Test Bank and Resource Guide (which includes a correlation to the AP\* Statistics course standards), Audio Podcasts, PowerPoint slides, and Graphing Calculator Manual. A replacement CD is available for purchase. (ISBN-13: 978-0-13-136349-6; ISBN-10: 0-13-136349-2)

**Companion DVD** A multimedia program on DVD designed to support learning chapter by chapter comes with the Teacher's Edition. It may be purchased separately for individual students or as a lab version (per work station). A replacement DVD is available for purchase. (ISBN-13: 978-0-13-136608-4; ISBN-10: 0-13-136608-4) The DVD holds a number of supporting materials, including:

- **ActivStats® for Data Desk.** The award-winning *ActivStats* multimedia program supports learning chapter by chapter with the book. It complements the book with videos of real-world stories, worked examples, animated expositions of each of the major Statistics topics, and tools for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes 17 short video clips; 170 animated activities and teaching applets; 300 data sets; 1,000 homework exercises, many with links to Data Desk files; interactive graphs, simulations, activities for the TI-Nspire graphing calculator, visualization tools, and much more.
- **Data Desk** statistics package.
- **TI-Nspire activities.** These investigations and demonstrations for the TI-Nspire handheld illustrate and explore important concepts from each chapter.
- **DDXL**, an Excel add-in, adds sound statistics and statistical graphics capabilities to Excel. DDXL adds, among other capabilities, boxplots, histograms, statistical scatterplots, normal probability plots, and statistical inference procedures not available in Excel's Data Analysis pack.
- **Data.** Data for exercises marked **T** are available on the DVD and at [www.aw.com/boc](http://www.aw.com/boc) formatted for Data Desk, Excel, JMP, MINITAB, SPSS, and the TI calculators, and as text files suitable for these and virtually any other statistics software.
- **Additional Chapters.** Two additional chapters cover **Analysis of Variance** (Chapter 28) and **Multiple Regression** (Chapter 29). These topics point the way to further study in Statistics.

**ActivStats®** The award-winning *ActivStats* multimedia program supports learning chapter by chapter with the book. It is available as a standalone DVD, or in a lab version (per work station). It complements the book with videos of real-world stories, worked examples, animated expositions of each of the major Statistics topics, and tools

for performing simulations, visualizing inference, and learning to use statistics software. The new version of *ActivStats* includes 17 short video clips; 170 animated activities and teaching applets; 300 data sets; 1,000 homework exercises, many with links to Data Desk files; interactive graphs, simulations, visualization tools, and much more. *ActivStats* (Mac and PC) is available in an all-in-one version for Data Desk, Excel, JMP, MINITAB, and SPSS. This DVD also includes Data Desk statistical software. For more information on options for purchasing *ActivStats*, contact Customer Service at 1-800-848-9500.

**MathXL® for School** is a powerful online homework, tutorial, and assessment system that accompanies Pearson textbooks in Statistics. With *MathXL for School*, teachers can create, edit, and assign online homework and tests using algorithmically generated exercises correlated at the objective level to the textbook. They can also create and assign their own online exercises and import TestGen tests for added flexibility. All student work is tracked in *MathXL for School's* online gradebook. Students can take chapter tests in *MathXL for School* and receive personalized study plans based on their test results. The study plan diagnoses weaknesses and links students directly to tutorial exercises for the objectives they need to study and retest. Students can also access supplemental animations directly from selected exercises. *MathXL for School* is available to qualified adopters. For more information, visit our Web site at [www.MathXLforSchool.com](http://www.MathXLforSchool.com), or contact your Pearson sales representative.

**StatCrunch** is a powerful online tool that provides an interactive environment for doing Statistics. *StatCrunch* can be used for both numerical and graphical data analysis, and uses interactive graphics to illustrate the connection between objects selected in a graph and the underlying data. *StatCrunch* may be purchased in a Registration Packet of 10 "redemptions." One redemption is for one student for 12 months beginning at the time of registration. Teacher access for *StatCrunch* adopters or for those wishing to preview the product may be obtained by filling out the form at [www.pearsonschool.com/access\\_request](http://www.pearsonschool.com/access_request) (ISBN-13: 978-0-13-136416-5; ISBN-10: 0-13-136416-2)

**Video Lectures on DVD with Subtitles** feature the textbook authors reviewing the high points of each chapter. The presentations continue the same student-friendly style and emphasis on critical thinking as the text. The DVD format makes it easy and convenient to watch the videos from a computer at home or on campus. (ISBN 13: 978-0-321-57103-8; ISBN-10: 0-321-57103-7)

**Companion Web Site ([www.aw.com/boc](http://www.aw.com/boc))** provides additional resources for instructors and students.



## Acknowledgments

Many people have contributed to this book in all three of its editions. This edition would have never seen the light of day without the assistance of the incredible team at Addison-Wesley. Our editor in chief, Deirdre Lynch, was central to the genesis, development, and realization of the book from day one. Chris Cummings, acquisitions editor, provided much needed support. Chere Bemelmans, senior project editor, kept us on task as much as humanly possible. Sheila Spinney, senior production supervisor, kept the cogs from getting into the wheels where they often wanted to wander. Christina Lepre, assistant editor, and Kathleen DeChavez, marketing assistant, were essential in managing all of the behind-the-scenes work that needed to be done. Christine Stavrou, media producer, put together a top-notch media package for this book. Barbara T. Atkinson, senior designer, and Geri Davis are responsible for the wonderful way the book looks. Carol Melville, manufacturing buyer, and Ginny Michaud, senior media buyer, worked miracles to get this book and DVD in your hands, and Greg Tobin, publisher, was supportive and good-humored throughout all aspects of the project. Special thanks go out to Pre-Press PMG, the compositor, for the wonderful work they did on this book, and in particular to Laura Hakala, senior project manager, for her close attention to detail. We'd also like to thank our accuracy checkers whose monumental task was to make sure we said what we thought we were saying. They are Jackie Miller, The Ohio State University; Douglas Cashing, St. Bonaventure University; Jared Derksen, Rancho Cucamonga High School; and Susan Blackwell, First Flight High School.

We extend our sincere thanks for the suggestions and contributions made by the following reviewers of this edition:

Allen Back, *Cornell University, New York*

Susan Blackwell, *First Flight High School, North Carolina*

Kevin Crowther, *Lake Orion High School, Michigan*

Sam Erickson, *North High School, Wisconsin*

Guillermo Leon, *Coral Reef High School, Florida*

Martha Lowther, *The Tatnall School, Delaware*

Karl Ronning, *Davis Senior High School, California*

Agatha Shaw, *Valencia Community College, Florida*

We extend our sincere thanks for the suggestions and contributions made by the following reviewers, focus group participants, and class-testers of the previous edition:

John Arko, *Glenbrook South High School, IL*

Kathleen Arthur, *Shaker High School, NY*

Beverly Beemer, *Ruben S. Ayala High School, CA*

Judy Bevington, *Santa Maria High School, CA*

Susan Blackwell, *First Flight High School, NC*

Gail Brooks, *McLennan Community College, TX*

Walter Brown, *Brackenridge High School, TX*

Darin Clift, *Memphis University School, TN*

Bill Craine, *Ithaca High School, NY*

Sybil Coley, *Woodward Academy, GA*

Caroline DiTullio, *Summit High School, NJ*

Jared Derksen, *Rancho Cucamonga High School, CA*

Laura Estersohn, *Scarsdale High School, NY*

Laura Favata, *Niskayuna High School, NY*

David Ferris, *Noblesville High School, IN*

Linda Gann, *Sandra Day O'Connor High School, TX*

Randall Groth, *Illinois State University, IL*

Donnie Hallstone, *Green River Community College, WA*

Howard W. Hand, *St. Marks School of Texas, TX*

Bill Hayes, *Foothill High School, CA*

Miles Hercamp, *New Palestine High School, IN*

Michelle Hipke, *Glen Burnie Senior High School, MD*

Carol Huss, *Independence High School, NC*

Sam Jovell, *Niskayuna High School, NY*

Peter Kaczmar, *Lower Merion High School, PA*  
 John Kotmel, *Lansing High School, NY*  
 Beth Lazerick, *St. Andrews School, FL*  
 Michael Legacy, *Greenhill School, TX*  
 John Lieb, *The Roxbury Latin School, MA*  
 John Maceli, *Ithaca College, NY*  
 Jim Miller, *Alta High School, UT*  
 Timothy E. Mitchell, *King Philip Regional High School, MA*

Maxine Nesbitt, *Carmel High School, IN*  
 Elizabeth Ann Przybysz, *Dr. Phillips High School, FL*  
 Diana Podhrasky, *Hillcrest High School, TX*  
 Rochelle Robert, *Nassau Community College, NY*  
 Bruce Saathoff, *Centennial High School, CA*  
 Murray Siegel, *Sam Houston State University, TX*  
 Chris Sollars, *Alamo Heights High School, TX*  
 Darren Starnes, *The Webb Schools, CA*

PART

I

# Exploring and Understanding Data

## Chapter 1

Stats Starts Here

## Chapter 2

Data

## Chapter 3

Displaying and Describing Categorical Data

## Chapter 4

Displaying and Summarizing  
Quantitative Data

## Chapter 5

Understanding and Comparing  
Distributions

## Chapter 6

The Standard Deviation  
as a Ruler and the  
Normal Model

Stats Starts Here<sup>1</sup>

*“But where shall I begin?”  
asked Alice. “Begin at the  
beginning,” the King said  
gravely, “and go on till you  
come to the end: then stop.”*

—Lewis Carroll,  
*Alice’s Adventures  
in Wonderland*

Statistics gets no respect. People say things like “You can prove anything with Statistics.” People will write off a claim based on data as “just a statistical trick.” And Statistics courses don’t have the reputation of being students’ first choice for a fun elective.

But Statistics *is* fun. That’s probably not what you heard on the street, but it’s true. Statistics is about how to think clearly with data. A little practice thinking statistically is all it takes to start seeing the world more clearly and accurately.

## So, What Is (Are?) Statistics?

Q: What is Statistics?

A: Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world.

Q: What are statistics?

A: Statistics (plural) are particular calculations made from data.

Q: So what is data?

A: You mean, “what *are* data?” Data is the plural form. The singular is datum.

Q: OK, OK, so what are data?

A: Data are values along with their context.

It seems every time we turn around, someone is collecting data on us, from every purchase we make in the grocery store, to every click of our mouse as we surf the Web. The United Parcel Service (UPS) tracks every package it ships from one place to another around the world and stores these records in a giant database. You can access part of it if you send or receive a UPS package. The database is about 17 terabytes big—about the same size as a database that contained every book in the Library of Congress would be. (But, we suspect, not *quite* as interesting.) What can anyone hope to do with all these data?

Statistics plays a role in making sense of the complex world in which we live today. Statisticians assess the risk of genetically engineered foods or of a new drug being considered by the Food and Drug Administration (FDA). They predict the number of new cases of AIDS by regions of the country or the number of customers likely to respond to a sale at the mall. And statisticians help scientists and social scientists understand how unemployment is related to environmental controls, whether enriched early education af-

<sup>1</sup> This chapter might have been called “Introduction,” but nobody reads the introduction, and we wanted you to read this. We feel safe admitting this here, in the footnote, because nobody reads footnotes either.

The ads say, "Don't drink and drive; you don't want to be a statistic." But you can't be a statistic.

We say: "Don't be a datum."

fects later performance of school children, and whether vitamin C really prevents illness. Whenever there are data and a need for understanding the world, you need Statistics.

So our objectives in this book are to help you develop the insights to think clearly about the questions, use the tools to show what the data are saying, and acquire the skills to tell clearly what it all means.



FRAZZ reprinted by permission of United Feature Syndicate, Inc.

## Statistics in a Word

Statistics is about variation.

Data vary because we don't see everything and because even what we do see and measure, we measure imperfectly.

So, in a very basic way, Statistics is about the real, imperfect world in which we live.

It can be fun, and sometimes useful, to summarize a discipline in only a few words. So,

Economics is about . . . *Money (and why it is good).*

Psychology: *Why we think what we think (we think).*

Biology: *Life.*

Anthropology: *Who?*

History: *What, where, and when?*

Philosophy: *Why?*

Engineering: *How?*

Accounting: *How much?*

In such a caricature, Statistics is about . . . **Variation.**

Data vary. People are different. We can't see everything, let alone measure it all. And even what we do measure, we measure imperfectly. So the data we wind up looking at and basing our decisions on provide, at best, an imperfect picture of the world. This fact lies at the heart of what Statistics is all about. How to make sense of it is a central challenge of Statistics.

## So, How Will This Book Help?

A fair question. Most likely, this book will not turn out to be quite what you expected.

What's different?

*Close your eyes and open the book to a page at random. Is there a graph or table on that page? Do that again, say, 10 times. We'll bet you saw data displayed in many ways, even near the back of the book and in the exercises.*

We can better understand everything we do with data by making pictures. This book leads you through the entire process of thinking about a problem, finding and showing results, and telling others about what you have discovered. At each of these steps, we display data for better understanding and insight.

You looked at only a few randomly selected pages to get an impression of the entire book. We'll see soon that doing so was sound Statistics practice and reasoning.

*Next, pick a chapter and read the first two sentences. (Go ahead; we'll wait.)*

We'll bet you didn't see anything about Statistics. Why? Because the best way to understand Statistics is to see it at work. In this book, chapters usually start by presenting a story and posing questions. That's when Statistics really gets down to work.

There are three simple steps to doing Statistics right: *think, show, and tell*:



**Think** first. Know where you're headed and why. It will save you a lot of work.



**Show** is what most folks think Statistics is about. The *mechanics* of calculating statistics and making displays is important, but not the most important part of Statistics.



**Tell** what you've learned. Until you've explained your results so that someone else can understand your conclusions, the job is not done.

**FOR EXAMPLE**

**STEP-BY-STEP**

The best way to learn new skills is to take them out for a spin. In **For Example** boxes you'll see brief ways to apply new ideas and methods as you learn them. You'll also find more comprehensive worked examples called **Step-by-Steps**. These show you fully worked solutions side by side with commentary and discussion, modeling the way statisticians attack and solve problems. They illustrate how to think about the problem, what to show, and how to tell what it all means. These step-by-step examples will show you how to produce the kind of solutions instructors hope to see.

Sometimes, in the middle of the chapter, we've put a section called **Just Checking** . . . There you'll find a few short questions you can answer without much calculation—a quick way to check to see if you've understood the basic ideas in the chapter. You'll find the answers at the end of the chapter's exercises.



## MATH BOX

Knowing where the formulas and procedures of Statistics come from and why they work will help you understand the important concepts. We'll provide brief, clear explanations of the mathematics that supports many of the statistical methods in **Math Boxes** like this.

## TI Tips

### Do statistics on your calculator!

How do I use  
this thing?

Although we'll show you all the formulas you need to understand the calculations, you will most often use a calculator or computer to perform the mechanics of a statistics problem. Your graphing calculator has a specialized program called a "statistics package." Each chapter contains **TI Tips** that teach you how to use it (and avoid doing most of the messy calculations).

**A S** If you have the DVD, you'll find **ActivStats** parallels the chapters in this book and includes expanded lessons and activities to increase your understanding of the material covered in the text.

TI-Nspire

*"Get your facts first, and then you can distort them as much as you please. (Facts are stubborn, but statistics are more pliable.)"*

—Mark Twain



From time to time, you'll see an icon like this in the margin to signal that the *ActivStats* multimedia materials on the available DVD in the back of the book have an activity that you might find helpful at this point. Typically, we've flagged simulations and interactive activities because they're the most fun and will probably help you see how things work best. The chapters in *ActivStats* are the same as those in the text—just look for the named activity in the corresponding chapter.

If you are using TI-Nspire™ technology, these margin icons will alert you to activities and demonstrations that can help you understand important ideas in the text. If you have the DVD that's available with this book, you'll find these there; if not, they're also available on the book's Web site [www.aw.com/bock](http://www.aw.com/bock).

One of the interesting challenges of Statistics is that, unlike in some math and science courses, there can be more than one right answer. This is why two statisticians can testify honestly on opposite sides of a court case. And it's why some people think that you can prove anything with statistics. But that's not true. People make mistakes using statistics, sometimes on purpose in order to mislead others. Most of the unintentional mistakes people make, though, are avoidable. We're not talking about arithmetic. More often, the mistakes come from using a method in the wrong situation or misinterpreting the results. Each chapter has a section called **What Can Go Wrong?** to help you avoid some of the most common mistakes.

**Time out.** From time to time, we'll take time out to discuss an interesting or important side issue. We indicate these by setting them apart like this.<sup>2</sup>

**A S** Introduction to (Your **Statistics Package**). *ActivStats* launches your statistics package (such as Data Desk) automatically. If you have the DVD, try it now.

#### ON THE COMPUTER

You'll find all sorts of stuff in margin notes, such as stories and quotations. For example:

*"Computers are useless. They can only give you answers."*

—Pablo Picasso

While Picasso underestimated the value of good statistics software, he did know that creating a solution requires more than just *Showing* an answer—it means you have to *Think* and *Tell*, too!

There are a number of statistics packages available for computers, and they differ widely in the details of how to use them and in how they present their results. But they all work from the same basic information and find the same results. Rather than adopt one package for this book, we present generic output and point out common features that you should look for. The . . . **on the Computer** section of most chapters (just before the exercises) holds this information. We also give a table of instructions to get you started on any of several commonly used packages, organized by chapters in Appendix B's Guide to Statistical Software.

At the end of each chapter, you'll see a brief summary of the important concepts you've covered in a section called **What Have We Learned?** That section includes a list of the **Terms** and a summary of the important **Skills** you've acquired in the chapter. You won't be able to learn the material from these summaries, but you can use them to check your knowledge of the important ideas in the chapter. If you have the skills, know the terms, and understand the concepts, you should be well prepared for the exam—and ready to use Statistics!

Beware: No one can learn Statistics just by reading or listening. The only way to learn it is to do it. So, of course, at the end of each chapter (except this one) you'll find **Exercises** designed to help you learn to use the Statistics you've just read about.

**T** Some exercises are marked with an orange **T**. You'll find the data for these exercises on the DVD in the back of the book or on the book's Web site at [www.aw.com/bock](http://www.aw.com/bock).

<sup>2</sup> Or in a footnote.

*“Far too many scientists have only a shaky grasp of the statistical techniques they are using. They employ them as an amateur chef employs a cookbook, believing the recipes will work without understanding why. A more cordon bleu attitude . . . might lead to fewer statistical soufflés failing to rise.”*

—*The Economist*, June 3, 2004, “**Sloppy stats shame science**”

We’ve paired up the exercises, putting similar ones together. So, if you’re having trouble doing an exercise, you will find a similar one either just before or just after it. You’ll find answers to the odd-numbered exercises at the back of the book. But these are only “answers” and not complete “solutions.” Huh? What’s the difference? The answers are sketches of the complete solutions. For most problems, your solution should follow the model of the Step-By-Step Examples. If your calculations match the numerical parts of the “answer” and your argument contains the elements shown in the answer, you’re on the right track. Your complete solution should explain the context, show your reasoning and calculations, and state your conclusions. Don’t fret too much if your numbers don’t match the printed answers to every decimal place. Statistics is more about getting the reasoning correct—pay more attention to how you interpret a result than what the digit in the third decimal place was.

In the real world, problems don’t come with chapters attached. So, in addition to the exercises at the ends of chapters, we’ve also collected a variety of problems at the end of each part of the text to make it more like the real world. This should help you to see whether you can sort out which methods to use when. If you can do that successfully, then you’ll know you understand Statistics.

## Onward!

It’s only fair to warn you: You can’t get there by just picking out the highlighted sentences and the summaries. This book is different. It’s not about memorizing definitions and learning equations. It’s deeper than that. And much more fun. But . . .

*You have to read the book!*<sup>3</sup>

---

<sup>3</sup> So, turn the page.





Many years ago, most stores in small towns knew their customers personally. If you walked into the hobby shop, the owner might tell you about a new bridge that had come in for your Lionel train set. The tailor knew your dad's size, and the hairdresser knew how your mom liked her hair. There are still some stores like that around today, but we're increasingly likely to shop at large stores, by phone, or on the Internet. Even so, when you phone an 800 number to buy new running shoes, customer service representatives may call you by your first name or ask about the socks you bought 6 weeks ago. Or the company may send an e-mail in October offering new head warmers for winter running. This company has millions of customers, and you called without identifying yourself. How did the sales rep know who you are, where you live, and what you had bought?

The answer is data. Collecting data on their customers, transactions, and sales lets companies track their inventory and helps them predict what their customers prefer. These data can help them predict what their customers may buy in the future so they know how much of each item to stock. The store can use the data and what it learns from the data to improve customer service, mimicking the kind of personal attention a shopper had 50 years ago.

Amazon.com opened for business in July 1995, billing itself as "Earth's Biggest Bookstore." By 1997, Amazon had a catalog of more than 2.5 million book titles and had sold books to more than 1.5 million customers in 150 countries. In 2006, the company's revenue reached \$10.7 billion. Amazon has expanded into selling a wide selection of merchandise, from \$400,000 necklaces<sup>1</sup> to yak cheese from Tibet to the largest book in the world.

Amazon is constantly monitoring and evolving its Web site to serve its customers better and maximize sales performance. To decide which changes to make to the site, the company experiments, collecting data and analyzing what works best. When you visit the Amazon Web site, you may encounter a different look or different suggestions and offers. Amazon statisticians want to know whether you'll follow the links offered, purchase the items suggested, or even spend a

*"Data is king at Amazon. Clickstream and purchase data are the crown jewels at Amazon. They help us build features to personalize the Web site experience."*

—Ronny Kohavi,  
Director of Data Mining  
and Personalization,  
Amazon.com



<sup>1</sup> Please get credit card approval before purchasing online.

longer time browsing the site. As Ronny Kohavi, director of Data Mining and Personalization, said, “Data trumps intuition. Instead of using our intuition, we experiment on the live site and let our customers tell us what works for them.”

## But What Are Data?

### THE W’S:

**WHO**

**WHAT**

and in what units

**WHEN**

**WHERE**

**WHY**

**HOW**

We bet you thought you knew this instinctively. Think about it for a minute. What exactly *do* we mean by “data”?

Do data have to be numbers? The amount of your last purchase in dollars is numerical data, but some data record names or other labels. The names in Amazon.com’s database are data, but not numerical.

Sometimes, data can have values that look like numerical values but are just numerals serving as labels. This can be confusing. For example, the ASIN (Amazon Standard Item Number) of a book, like 0321570448, may have a numerical value, but it’s really just another name for *Stats: Modeling the World*.

Data values, no matter what kind, are useless without their context. Newspaper journalists know that the lead paragraph of a good story should establish the “Five W’s”: *Who*, *What*, *When*, *Where*, and (if possible) *Why*. Often we add *How* to the list as well. Answering these questions can provide the **context** for data values. The answers to the first two questions are essential. If you can’t answer *Who* and *What*, you don’t have **data**, and you don’t have any useful information.

## Data Tables

Here are some data Amazon might collect:

B000001OAA	10.99	Chris G.	902	15783947	15.98	Kansas	Illinois	Boston
Canada	Samuel P.	Orange County	N	B000068ZVQ	Bad Blood	Nashville	Katherine H.	N
Mammals	10783489	Ohio	N	Chicago	12837593	11.99	Massachusetts	16.99
312	Monique D.	10675489	413	B0000015Y6	440	B000002BK9	Let Go	Y

**A S** **Activity: What Is (Are) Data?** Do you really know what’s data and what’s just numbers?

Try to guess what they represent. Why is that hard? Because these data have no *context*. If we don’t know *Who* they’re about or *What* they measure, these values are meaningless. We can make the meaning clear if we organize the values into a **data table** such as this one:

Purchase Order	Name	Ship to State/Country	Price	Area Code	Previous CD Purchase	Gift?	ASIN	Artist
10675489	Katharine H.	Ohio	10.99	440	Nashville	N	B0000015Y6	Kansas
10783489	Samuel P.	Illinois	16.99	312	Orange County	Y	B000002BK9	Boston
12837593	Chris G.	Massachusetts	15.98	413	Bad Blood	N	B000068ZVQ	Chicago
15783947	Monique D.	Canada	11.99	902	Let Go	N	B000001OAA	Mammals

Now we can see that these are four purchase records, relating to CD orders from Amazon. The column titles tell *What* has been recorded. The rows tell us *Who*. But be careful. Look at all the variables to see *Who* the variables are about. Even if people are involved, they may not be the *Who* of the data. For example, the *Who* here are the purchase orders (not the people who made the purchases).

A common place to find the *Who* of the table is the leftmost column. The other *W*'s might have to come from the company's database administrator.<sup>2</sup>

## Who

In general, the rows of a data table correspond to individual **cases** about *Whom* (or about which—if they're not people) we record some characteristics. These cases go by different names, depending on the situation. Individuals who answer a survey are referred to as *respondents*. People on whom we experiment are *subjects* or (in an attempt to acknowledge the importance of their role in the experiment) *participants*, but animals, plants, Web sites, and other inanimate subjects are often just called *experimental units*. In a database, rows are called *records*—in this example, purchase records. Perhaps the most generic term is **cases**. In the Amazon table, the cases are the individual CD orders.

Sometimes people just refer to data values as *observations*, without being clear about the *Who*. Be sure you know the *Who* of the data, or you may not know what the data say.

Often, the cases are a **sample** of cases selected from some larger **population** that we'd like to understand. Amazon certainly cares about its customers, but also wants to know how to attract all those other Internet users who may never have made a purchase from Amazon's site. To be able to generalize from the sample of cases to the larger population, we'll want the sample to be *representative* of that population—a kind of snapshot image of the larger world.

**A S** **Activity: Consider the Context** . . . Can you tell who's *Who* and what's *What*? And *Why*? This activity offers real-world examples to help you practice identifying the context.

### FOR EXAMPLE

#### Identifying the "Who"

In March 2007, *Consumer Reports* published an evaluation of large-screen, high-definition television sets (HDTVs). The magazine purchased and tested 98 different models from a variety of manufacturers.

**Question:** Describe the population of interest, the sample, and the *Who* of this study.

The magazine is interested in the performance of all HDTVs currently being offered for sale. It tested a sample of 98 sets, the "Who" for these data. Each HDTV set represents all similar sets offered by that manufacturer.

## What and Why

The characteristics recorded about each individual are called **variables**. These are usually shown as the columns of a data table, and they should have a name that identifies *What* has been measured. Variables may seem simple, but to really understand your variables, you must *Think* about what you want to know.

Although area codes are numbers, do we use them that way? Is 610 twice 305? Of course it is, but is that the question? Why would we want to know whether Allentown, PA (area code 610), is twice Key West, FL (305)? Variables play different roles, and you can't tell a variable's role just by looking at it.

Some variables just tell us what group or category each individual belongs to. Are you male or female? Pierced or not? . . . What kinds of things can we learn about variables like these? A natural start is to *count* how many cases belong in each category. (Are you listening to music while reading this? We could count

<sup>2</sup> In database management, this kind of information is called "metadata."

It is wise to be careful. The *What* and *Why* of area codes are not as simple as they may first seem. When area codes were first introduced, AT&T was still the source of all telephone equipment, and phones had dials.



To reduce wear and tear on the dials, the area codes with the lowest digits (for which the dial would have to spin least) were assigned to the most populous regions—those with the most phone numbers and thus the area codes most likely to be dialed. New York City was assigned 212, Chicago 312, and Los Angeles 213, but rural upstate New York was given 607, Joliet was 815, and San Diego 619. For that reason, at one time the numerical value of an area code could be used to guess something about the population of its region. Now that phones have push-buttons, area codes have finally become just categories.

By international agreement, the International System of Units links together all systems of weights and measures. There are seven base units from which all other physical units are derived:

- |                       |          |
|-----------------------|----------|
| • Distance            | Meter    |
| • Mass                | Kilogram |
| • Time                | Second   |
| • Electric current    | Ampere   |
| • Temperature         | °Kelvin  |
| • Amount of substance | Mole     |
| • Intensity of light  | Candela  |

**AS** **Activity: Recognize variables measured in a variety of ways.** This activity shows examples of the many ways to measure data.

**AS** **Activities: Variables.** Several activities show you how to begin working with data in your statistics package.

the number of students in the class who were and the number who weren't.) We'll look for ways to compare and contrast the sizes of such categories.

Some variables have measurement **units**. Units tell how each value has been measured. But, more importantly, units such as yen, cubits, carats, angstroms, nanoseconds, miles per hour, or degrees Celsius tell us the *scale* of measurement. The units tell us how much of something we have or how far apart two values are. Without units, the values of a measured variable have no meaning. It does little good to be promised a raise of 5000 a year if you don't know whether it will be paid in euros, dollars, yen, or Estonian krooni.

What kinds of things can we learn about measured variables? We can do a lot more than just counting categories. We can look for patterns and trends. (How much did you pay for your last movie ticket? What is the range of ticket prices available in your town? How has the price of a ticket changed over the past 20 years?)

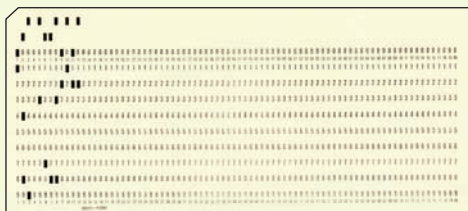
When a variable names categories and answers questions about how cases fall into those categories, we call it a **categorical variable**.<sup>3</sup> When a measured variable with units answers questions about the quantity of what is measured, we call it a **quantitative variable**. These types can help us decide what to do with a variable, but they are really more about what we hope to learn from a variable than about the variable itself. It's the questions we ask a variable (the *Why* of our analysis) that shape how we think about it and how we treat it.

Some variables can answer questions only about categories. If the values of a variable are words rather than numbers, it's a good bet that it is categorical. But some variables can answer both kinds of questions. Amazon could ask for your *Age* in years. That seems quantitative, and would be if the company wanted to know the average age of those customers who visit their site after 3 a.m. But suppose Amazon wants to decide which CD to offer you in a special deal—one by Raffi, Blink-182, Carly Simon, or Mantovani—and needs to be sure to have adequate supplies on hand to meet the demand. Then thinking of your age in one of the categories—child, teen, adult, or senior—might be more useful. If it isn't clear whether a variable is categorical or quantitative, think about *Why* you are looking at it and what you want it to tell you.

A typical course evaluation survey asks, "How valuable do you think this course will be to you?": 1 = Worthless; 2 = Slightly; 3 = Middling; 4 = Reasonably; 5 = Invaluable. Is *Educational Value* categorical or quantitative? Once again, we'll look to the *Why*. A teacher might just count the number of students who gave each response for her course, treating *Educational Value* as a categorical variable. When she wants to see whether the course is improving, she might treat the responses as the *amount* of perceived value—in effect, treating the variable as quantitative. But what are the units? There is certainly an *order* of perceived worth: Higher numbers indicate higher perceived worth. A course that averages 4.5 seems more valuable than one that averages 2, but we should be careful about treating *Educational Value* as

<sup>3</sup> You may also see it called a *qualitative variable*.

One tradition that hangs on in some quarters is to name variables with cryptic abbreviations written in uppercase letters. This can be traced back to the 1960s, when the very first statistics computer programs were controlled with instructions punched on cards. The earliest punch card equipment used only uppercase letters, and the earliest statistics programs limited variable names to six or eight characters, so variables were called things like PRSRF3. Modern programs do not have such restrictive limits, so there is no reason for variable names that you wouldn't use in an ordinary sentence.



purely quantitative. To treat it as quantitative, she'll have to imagine that it has "educational value units" or some similar arbitrary construction. Because there are no natural units, she should be cautious. Variables like this that report order without natural units are often called "ordinal" variables. But saying "that's an ordinal variable" doesn't get you off the hook. You must still look to the *Why* of your study to decide whether to treat it as categorical or quantitative.

## FOR EXAMPLE

### Identifying "What" and "Why" of HDTVs.

**Recap:** A *Consumer Reports* article about 98 HDTVs lists each set's manufacturer, cost, screen size, type (LCD, plasma, or rear projection), and overall performance score (0–100).

**Question:** Are these variables categorical or quantitative? Include units where appropriate, and describe the "Why" of this investigation.

The "what" of this article includes the following variables:

- manufacturer (categorical);
- cost (in dollars, quantitative);
- screen size (in inches, quantitative);
- type (categorical);
- performance score (quantitative).

The magazine hopes to help consumers pick a good HDTV set.

## Counts Count

In Statistics, we often count things. When Amazon considers a special offer of free shipping to customers, it might first analyze how purchases are shipped. They'd probably start by counting the number of purchases shipped by ground transportation, by second-day air, and by overnight air. Counting is a natural way to summarize the categorical variable *Shipping Method*. So every time we see counts, does that mean the variable is categorical? Actually, no.

We also use counts to measure the amounts of things. How many songs are on your digital music player? How many classes are you taking this semester? To measure these quantities, we'd naturally count. The variables (*Songs*, *Classes*) would be quantitative, and we'd consider the units to be "number of . . ." or, generically, just "counts" for short.

So we use counts in two different ways. When we count the cases in each category of a categorical variable, the category labels are the *What* and the individuals counted are the *Who* of our data. The counts themselves are not the

**AS** **Activity: Collect data in an experiment on yourself.** With the computer, you can experiment on yourself and then save the data. Go on to the subsequent related activities to check your understanding.

data, but are something we summarize about the data. Amazon counts the number of purchases in each category of the categorical variable *Shipping Method*. For this purpose (the *Why*), the *What* is shipping method and the *Who* is purchases.

Shipping Method	Number of Purchases
Ground	20,345
Second-day	7,890
Overnight	5,432

Other times our focus is on the amount of something, which we measure by counting. Amazon might record the number of teenage customers visiting their site each month to track customer growth and forecast CD sales (the *Why*). Now the *What* is *Teens*, the *Who* is *Months*, and the units are *Number of Teenage Customers*. *Teen* was a category when we looked at the categorical variable *Age*. But now it is a quantitative variable in its own right whose amount is measured by counting the number of customers.

Month	Number of Teenage Customers
January	123,456
February	234,567
March	345,678
April	456,789
May	...
...	...

## Identifying Identifiers

What's your student ID number? It is numerical, but is it a quantitative variable? No, it doesn't have units. Is it categorical? Yes, but it is a special kind. Look at how many categories there are and at how many individuals are in each. There are as many categories as individuals and only one individual in each category. While it's easy to count the totals for each category, it's not very interesting. Amazon wants to know who you are when you sign in again and doesn't want to confuse you with some other customer. So it assigns you a unique identifier.

Identifier variables themselves don't tell us anything useful about the categories because we know there is exactly one individual in each. However, they are crucial in this age of large data sets. They make it possible to combine data from different sources, to protect confidentiality, and to provide unique labels. The variables *UPS Tracking Number*, *Social Security Number*, and Amazon's *ASIN* are all examples of identifier variables.

You'll want to recognize when a variable is playing the role of an identifier so you won't be tempted to analyze it. There's probably a list of unique ID numbers for students in a class (so they'll each get their own grade confidentially), but you might worry about the professor who keeps track of the average of these numbers from class to class. Even though this year's average ID number happens to be higher than last's, it doesn't mean that the students are better.

## Where, When, and How

**AS**

**Self-Test: Review concepts about data.** Like the Just Checking sections of this textbook, but interactive. (Usually, we won't reference the *ActivStats* self-tests here, but look for one whenever you'd like to check your understanding or review material.)

We must know *Who*, *What*, and *Why* to analyze data. Without knowing these three, we don't have enough to start. Of course, we'd always like to know more. The more we know about the data, the more we'll understand about the world.

If possible, we'd like to know the **When** and **Where** of data as well. Values recorded in 1803 may mean something different than similar values recorded last year. Values measured in Tanzania may differ in meaning from similar measurements made in Mexico.

**How** the data are collected can make the difference between insight and nonsense. As we'll see later, data that come from a voluntary survey on the Internet are almost always worthless. One primary concern of Statistics, to be discussed in Part III, is the design of sound methods for collecting data.

Throughout this book, whenever we introduce data, we'll provide a margin note listing the W's (and H) of the data. It's a habit we recommend. The first step of any data analysis is to know why you are examining the data (what you want to know), whom each row of your data table refers to, and what the variables (the columns of the table) record. These are the *Why*, the *Who*, and the *What*. Identifying them is a key part of the *Think* step of any analysis. Make sure you know all three before you proceed to *Show* or *Tell* anything about the data.



### JUST CHECKING

In the 2003 Tour de France, Lance Armstrong averaged 40.94 kilometers per hour (km/h) for the entire course, making it the fastest Tour de France in its 100-year history. In 2004, he made history again by winning the race for an unprecedented sixth time. In 2005, he became the only 7-time winner and once again set a new record for the fastest average speed. You can find data on all the Tour de France races on the DVD. Here are the first three and last ten lines of the data set. Keep in mind that the entire data set has nearly 100 entries.

- List as many of the W's as you can for this data set.
- Classify each variable as categorical or quantitative; if quantitative, identify the units.



Year	Winner	Country of origin	Total time (h/min/s)	Avg. speed (km/h)	Stages	Total distance ridden (km)	Starting riders	Finishing riders
1903	Maurice Garin	France	94.33.00	25.3	6	2428	60	21
1904	Henri Cornet	France	96.05.00	24.3	6	2388	88	23
1905	Louis Trousselier	France	112.18.09	27.3	11	2975	60	24
⋮								
1999	Lance Armstrong	USA	91.32.16	40.30	20	3687	180	141
2000	Lance Armstrong	USA	92.33.08	39.56	21	3662	180	128
2001	Lance Armstrong	USA	86.17.28	40.02	20	3453	189	144
2002	Lance Armstrong	USA	82.05.12	39.93	20	3278	189	153
2003	Lance Armstrong	USA	83.41.12	40.94	20	3427	189	147
2004	Lance Armstrong	USA	83.36.02	40.53	20	3391	188	147
2005	Lance Armstrong	USA	86.15.02	41.65	21	3608	189	155
2006	Óscar Periero	Spain	89.40.27	40.78	20	3657	176	139
2007	Alberto Contador	Spain	91.00.26	38.97	20	3547	189	141
2008	Carlos Sastre	Spain	87.52.52	40.50	21	3559	199	145

**There's a world of data on the Internet.** These days, one of the richest sources of data is the Internet. With a bit of practice, you can learn to find data on almost any subject. Many of the data sets we use in this book were found in this way. The Internet has both advantages and disadvantages as a source of data. Among the advantages are the fact that often you'll be able to find even more current data than those we present. The disadvantage is that references to Internet addresses can "break" as sites evolve, move, and die.

Our solution to these challenges is to offer the best advice we can to help you search for the data, wherever they may be residing. We usually point you to a Web site. We'll sometimes suggest search terms and offer other guidance.

Some words of caution, though: Data found on Internet sites may not be formatted in the best way for use in statistics software. Although you may see a data table in standard form, an attempt to copy the data may leave you with a single column of values. You may have to work in your favorite statistics or spreadsheet program to reformat the data into variables. You will also probably want to remove commas from large numbers and such extra symbols as money indicators (\$, ¥, £); few statistics packages can handle these.

## WHAT CAN GO WRONG?

- ▶ **Don't label a variable as categorical or quantitative without thinking about the question you want it to answer.** The same variable can sometimes take on different roles.
- ▶ **Just because your variable's values are numbers, don't assume that it's quantitative.** Categories are often given numerical labels. Don't let that fool you into thinking they have quantitative meaning. Look at the context.
- ▶ **Always be skeptical.** One reason to analyze data is to discover the truth. Even when you are told a context for the data, it may turn out that the truth is a bit (or even a lot) different. The context colors our interpretation of the data, so those who want to influence what you think may slant the context. A survey that seems to be about all students may in fact report just the opinions of those who visited a fan Web site. The question that respondents answered may have been posed in a way that influenced their responses.

### TI Tips

## Working with data

You'll need to be able to enter and edit data in your calculator. Here's how.

### To enter data:

Hit the **STAT** button, and choose **EDIT** from the menu. You'll see a set of columns labeled **L1**, **L2**, and so on. Here is where you can enter, change, or delete a set of data.

Let's enter the heights (in inches) of the five starting players on a basketball team: 71, 75, 75, 76, and 80. Move the cursor to the space under **L1**, type in 71, and hit **ENTER** (or the down arrow). There's the first player. Now enter the data for the rest of the team.

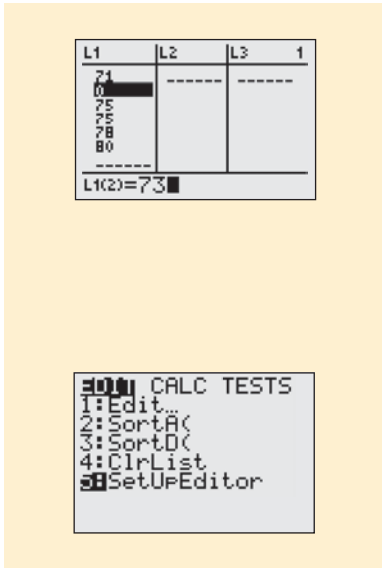
### To change a datum:

Suppose the 76" player grew since last season; his height should be listed as 78". Use the arrow keys to move the cursor onto the 76, then change the value and **ENTER** the correction.

L1	L2	L3	1
71	-----	-----	
75			
75			
76			
80			
L1(6)=			

L1	L2	L3	1
71	-----	-----	
75			
75			
78			
80			
-----			
L1(4)=78			





**To add more data:**

We want to include the sixth man, 73" tall. It would be easy to simply add this new datum to the end of the list. However, sometimes the order of the data matters, so let's place this datum in numerical order. Move the cursor to the desired position (atop the first 75). Hit **2ND INS**, then **ENTER** the 73 in the new space.

**To delete a datum:**

The 78" player just quit the team. Move the cursor there. Hit **DEL**. Bye.

**To clear the datalist:**

Finished playing basketball? Move the cursor atop the **L1**. Hit **CLEAR**, then **ENTER** (or down arrow). You should now have a blank datalist, ready for you to enter your next set of values.

**Lost a datalist?**

Oops! Is **L1** now missing entirely? Did you delete **L1** by mistake, instead of just *clearing* it? Easy problem to fix: buy a new calculator. No? OK, then simply go to the **STAT EDIT** menu, and run **SetUpEditor** to recreate all the lists.



## WHAT HAVE WE LEARNED?

We've learned that data are information in a context.

- ▶ The W's help nail down the context: *Who, What, Why, Where, When, and how*.
- ▶ We must know at least the *Who, What, and Why* to be able to say anything useful based on the data. The *Who* are the cases. The *What* are the *variables*. A variable gives information about each of the cases. The *Why* helps us decide which way to treat the variables.

We treat variables in two basic ways: as *categorical* or *quantitative*.

- ▶ Categorical variables identify a category for each case. Usually, we think about the counts of cases that fall into each category. (An exception is an identifier variable that just names each case.)
- ▶ Quantitative variables record measurements or amounts of something; they must have *units*.
- ▶ Sometimes we treat a variable as categorical or quantitative depending on what we want to learn from it, which means that some variables can't be pigeonholed as one type or the other. That's an early hint that in Statistics we can't always pin things down precisely.

### Terms

Context	8. The context ideally tells <i>Who</i> was measured, <i>What</i> was measured, <i>How</i> the data were collected, <i>Where</i> the data were collected, and <i>When</i> and <i>Why</i> the study was performed.
Data	8. Systematically recorded information, whether numbers or labels, together with its context.
Data table	8. An arrangement of data in which each row represents a case and each column represents a variable.
Case	9. A case is an individual about whom or which we have data.
Population	9. All the cases we wish we knew about.
Sample	9. The cases we actually examine in seeking to understand the much larger population.
Variable	9. A variable holds information about the same characteristic for many cases.
Units	10. A quantity or amount adopted as a standard of measurement, such as dollars, hours, or grams.
Categorical variable	10. A variable that names categories (whether with words or numerals) is called categorical.
Quantitative variable	10. A variable in which the numbers act as numerical values is called quantitative. Quantitative variables always have units.

## Skills



- ▶ Be able to identify the *Who*, *What*, *When*, *Where*, *Why*, and *How* of data, or recognize when some of this information has not been provided.
- ▶ Be able to identify the cases and variables in any data set.
- ▶ Be able to identify the population from which a sample was chosen.
- ▶ Be able to classify a variable as categorical or quantitative, depending on its use.
- ▶ For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).



- ▶ Be able to describe a variable in terms of its *Who*, *What*, *When*, *Where*, *Why*, and *How* (and be prepared to remark when that information is not provided).

## DATA ON THE COMPUTER

**A S**
**Activity: Examine the**

**Data.** Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- ▶ Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- ▶ Where to put the data. (Usually this is handled automatically.)
- ▶ What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

## EXERCISES

1. **Voters.** A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?
  2. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?
  3. **Medicine.** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?
  4. **Stress.** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?
- (Exercises 5–12) For each description of data, identify *Who* and *What* were investigated and the population of interest.

## Skills



- ▶ Be able to identify the *Who*, *What*, *When*, *Where*, *Why*, and *How* of data, or recognize when some of this information has not been provided.
- ▶ Be able to identify the cases and variables in any data set.
- ▶ Be able to identify the population from which a sample was chosen.
- ▶ Be able to classify a variable as categorical or quantitative, depending on its use.
- ▶ For any quantitative variable, be able to identify the units in which the variable has been measured (or note that they have not been provided).



- ▶ Be able to describe a variable in terms of its *Who*, *What*, *When*, *Where*, *Why*, and *How* (and be prepared to remark when that information is not provided).

## DATA ON THE COMPUTER

**A S**
**Activity: Examine the**

**Data.** Take a look at your own data from your experiment (p. 12) and get comfortable with your statistics package as you find out about the experiment test results.

Most often we find statistics on a computer using a program, or *package*, designed for that purpose. There are many different statistics packages, but they all do essentially the same things. If you understand what the computer needs to know to do what you want and what it needs to show you in return, you can figure out the specific details of most packages pretty easily.

For example, to get your data into a computer statistics package, you need to tell the computer:

- ▶ Where to find the data. This usually means directing the computer to a file stored on your computer's disk or to data on a database. Or it might just mean that you have copied the data from a spreadsheet program or Internet site and it is currently on your computer's clipboard. Usually, the data should be in the form of a data table. Most computer statistics packages prefer the *delimiter* that marks the division between elements of a data table to be a *tab* character and the delimiter that marks the end of a case to be a *return* character.
- ▶ Where to put the data. (Usually this is handled automatically.)
- ▶ What to call the variables. Some data tables have variable names as the first row of the data, and often statistics packages can take the variable names from the first row automatically.

## EXERCISES

1. **Voters.** A February 2007 Gallup Poll question asked, "In politics, as of today, do you consider yourself a Republican, a Democrat, or an Independent?" The possible responses were "Democrat", "Republican", "Independent", "Other", and "No Response". What kind of variable is the response?
  2. **Mood.** A January 2007 Gallup Poll question asked, "In general, do you think things have gotten better or gotten worse in this country in the last five years?" Possible answers were "Better", "Worse", "No Change", "Don't Know", and "No Response". What kind of variable is the response?
  3. **Medicine.** A pharmaceutical company conducts an experiment in which a subject takes 100 mg of a substance orally. The researchers measure how many minutes it takes for half of the substance to exit the bloodstream. What kind of variable is the company studying?
  4. **Stress.** A medical researcher measures the increase in heart rate of patients under a stress test. What kind of variable is the researcher studying?
- (Exercises 5–12) For each description of data, identify *Who* and *What* were investigated and the population of interest.

5. **The news.** Find a newspaper or magazine article in which some data are reported. For the data discussed in the article, answer the questions above. Include a copy of the article with your report.
6. **The Internet.** Find an Internet source that reports on a study and describes the data. Print out the description and answer the questions above.
7. **Bicycle safety.** Ian Walker, a psychologist at the University of Bath, wondered whether drivers treat bicycle riders differently when they wear helmets. He rigged his bicycle with an ultrasonic sensor that could measure how close each car was that passed him. He then rode on alternating days with and without a helmet. Out of 2500 cars passing him, he found that when he wore his helmet, motorists passed 3.35 inches closer to him, on average, than when his head was bare. [*NY Times*, Dec. 10, 2006]
8. **Investments.** Some companies offer 401(k) retirement plans to employees, permitting them to shift part of their before-tax salaries into investments such as mutual funds. Employers typically match 50% of the employees' contribution up to about 6% of salary. One company, concerned with what it believed was a low employee participation rate in its 401(k) plan, sampled 30 other companies with similar plans and asked for their 401(k) participation rates.
9. **Honesty.** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University alternately taped two posters over the coffee station. During one week, it was a picture of flowers; during the other, it was a pair of staring eyes. They found that the average contribution was significantly higher when the eyes poster was up than when the flowers were there. Apparently, the mere feeling of being watched—even by eyes that were not real—was enough to encourage people to behave more honestly. [*NY Times*, Dec. 10, 2006]
10. **Movies.** Some motion pictures are profitable and others are not. Understandably, the movie industry would like to know what makes a movie successful. Data from 120 first-run movies released in 2005 suggest that longer movies actually make *less* profit.
11. **Fitness.** Are physically fit people less likely to die of cancer? An article in the May 2002 issue of *Medicine and Science in Sports and Exercise* reported results of a study that followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.
12. **Molten iron.** The Cleveland Casting Plant is a large, highly automated producer of gray and nodular iron automotive castings for Ford Motor Company. The company is interested in keeping the pouring temperature of the molten iron (in degrees Fahrenheit) close to the specified value of 2550 degrees. Cleveland Casting measured the pouring temperature for 10 randomly selected crankshafts.
 

(Exercises 13–26) For each description of data, identify the *W*'s, name the variables, specify for each variable whether its use indicates that it should be treated as categorical or quantitative, and, for any quantitative variable, identify the units in which it was measured (or note that they were not provided).
13. **Weighing bears.** Because of the difficulty of weighing a bear in the woods, researchers caught and measured 54 bears, recording their weight, neck size, length, and sex. They hoped to find a way to estimate weight from the other, more easily determined quantities.
14. **Schools.** The State Education Department requires local school districts to keep these records on all students: age, race or ethnicity, days absent, current grade level, standardized test scores in reading and mathematics, and any disabilities or special educational needs.
15. **Arby's menu.** A listing posted by the Arby's restaurant chain gives, for each of the sandwiches it sells, the type of meat in the sandwich, the number of calories, and the serving size in ounces. The data might be used to assess the nutritional value of the different sandwiches.
16. **Age and party.** The Gallup Poll conducted a representative telephone survey of 1180 American voters during the first quarter of 2007. Among the reported results were the voter's region (Northeast, South, etc.), age, party affiliation, and whether or not the person had voted in the 2006 midterm congressional election.
17. **Babies.** Medical researchers at a large city hospital investigating the impact of prenatal care on newborn health collected data from 882 births during 1998–2000. They kept track of the mother's age, the number of weeks the pregnancy lasted, the type of birth (cesarean, induced, natural), the level of prenatal care the mother had (none, minimal, adequate), the birth weight and sex of the baby, and whether the baby exhibited health problems (none, minor, major).
18. **Flowers.** In a study appearing in the journal *Science*, a research team reports that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years show that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.
19. **Herbal medicine.** Scientists at a major pharmaceutical firm conducted an experiment to study the effectiveness of an herbal compound to treat the common cold. They exposed each patient to a cold virus, then gave them either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed each patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of the benefits of the compound.
20. **Vineyards.** Business analysts hoping to provide information helpful to American grape growers compiled these data about vineyards: size (acres), number of years in existence, state, varieties of grapes grown, average case price, gross sales, and percent profit.

- 21. Streams.** In performing research for an ecology class, students at a college in upstate New York collect data on streams each year. They record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (limestone, shale, or mixed), the acidity of the water (pH), the temperature (°C), and the BCI (a numerical measure of biological diversity).
- 22. Fuel economy.** The Environmental Protection Agency (EPA) tracks fuel economy of automobiles based on information from the manufacturers (Ford, Toyota, etc.). Among the data the agency collects are the manufacturer, vehicle type (car, SUV, etc.), weight, horsepower, and gas mileage (mpg) for city and highway driving.
- 23. Refrigerators.** In 2006, *Consumer Reports* published an article evaluating refrigerators. It listed 41 models, giving the brand, cost, size (cu ft), type (such as top freezer), estimated annual energy cost, an overall rating (good, excellent, etc.), and the repair history for that brand (percentage requiring repairs over the past 5 years).

- 24. Walking in circles.** People who get lost in the desert, mountains, or woods often seem to wander in circles rather than walk in straight lines. To see whether people naturally walk in circles in the absence of visual clues, researcher Andrea Axtell tested 32 people on a football field. One at a time, they stood at the center of one goal line, were blindfolded, and then tried to walk to the other goal line. She recorded each individual's sex, height, handedness, the number of yards each was able to walk before going out of bounds, and whether each wandered off course to the left or the right. No one made it all the way to the far end of the field without crossing one of the sidelines. [STATS No. 39, Winter 2004]

- T 25. Horse race 2008.** The Kentucky Derby is a horse race that has been run every year since 1875 at Churchill Downs, Louisville, Kentucky. The race started as a 1.5-mile race, but in 1896, it was shortened to 1.25 miles because experts felt that 3-year-old horses shouldn't run such a long race that early in the season. (It has been run in May every year but one—1901—when it took place on April 29). Here are the data for the first four and several recent races.

Date	Winner	Margin (lengths)	Jockey	Winner's Payoff (\$)	Duration (min:sec)	Track Condition
May 17, 1875	Aristides	2	O. Lewis	2850	2:37.75	Fast
May 15, 1876	Vagrant	2	B. Swim	2950	2:38.25	Fast
May 22, 1877	Baden-Baden	2	W. Walker	3300	2:38.00	Fast
May 21, 1878	Day Star	1	J. Carter	4050	2:37.25	Dusty
.....						
May 1, 2004	Smarty Jones	2 3/4	S. Elliott	854800	2:04.06	Sloppy
May 7, 2005	Giacomo	1/2	M. Smith	5854800	2:02.75	Fast
May 6, 2006	Barbaro	6 1/2	E. Prado	1453200	2:01.36	Fast
May 5, 2007	Street Sense	2 1/4	C. Borel	1450000	2:02.17	Fast
May 3, 2008	Big Brown	4 3/4	K. Desormeaux	1451800	2:01.82	Fast

- T 26. Indy 2008.** The 2.5-mile Indianapolis Motor Speedway has been the home to a race on Memorial Day nearly every year since 1911. Even during the first race, there were controversies. Ralph Mulford was given the checkered flag first but took three extra laps just to make sure he'd completed 500 miles. When he finished, another driver, Ray Harroun, was being presented with the

winner's trophy, and Mulford's protests were ignored. Harroun averaged 74.6 mph for the 500 miles. In 2008, the winner, Scott Dixon, averaged 143.567 mph.

Here are the data for the first five races and five recent Indianapolis 500 races. Included also are the pole winners (the winners of the trial races, when each driver drives alone to determine the position on race day).

Year	Winner	Pole Position	Average Speed (mph)	Pole Winner	Average Pole Speed (mph)
1911	Ray Harroun	28	74.602	Lewis Strang	.
1912	Joe Dawson	7	78.719	Gil Anderson	.
1913	Jules Goux	7	75.933	Caleb Bragg	.
1914	René Thomas	15	82.474	Jean Chassagne	.
1915	Ralph DePalma	2	89.840	Howard Wilcox	98.580
...					
2004	Buddy Rice	1	138.518	Buddy Rice	220.024
2005	Dan Wheldon	16	157.603	Tony Kanaan	224.308
2006	Sam Hornish Jr.	1	157.085	Sam Hornish Jr.	228.985
2007	Dario Franchitti	3	151.744	Hélio Castroneves	225.817
2008	Scott Dixon	1	143.567	Scott Dixon	221.514



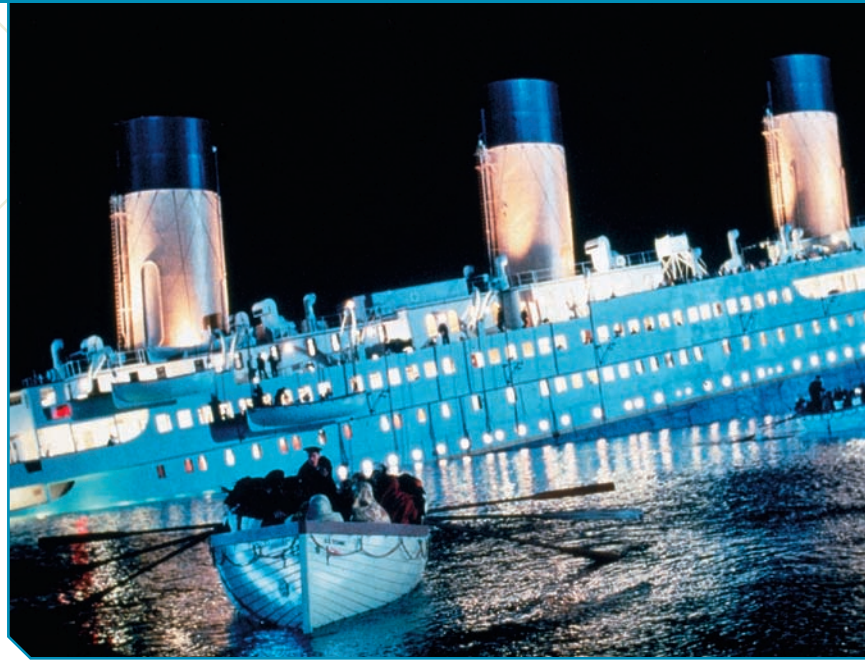
## JUST CHECKING Answers

1. Who—Tour de France races; What—year, winner, country of origin, total time, average speed, stages, total distance ridden, starting riders, finishing riders; How—official statistics at race; Where—France (for the most part); When—1903 to 2008; Why—not specified (To see progress in speeds of cycling racing?)

2.

Variable	Type	Units
Year	Quantitative or Categorical	Years
Winner	Categorical	
Country of Origin	Categorical	
Total Time	Quantitative	Hours/minutes/seconds
Average Speed	Quantitative	Kilometers per hour
Stages	Quantitative	Counts (stages)
Total Distance	Quantitative	Kilometers
Starting Riders	Quantitative	Counts (riders)
Finishing Riders	Quantitative	Counts (riders)

# Displaying and Describing Categorical Data



<b>WHO</b>	People on the <i>Titanic</i>
<b>WHAT</b>	Survival status, age, sex, ticket class
<b>WHEN</b>	April 14, 1912
<b>WHERE</b>	North Atlantic
<b>HOW</b>	A variety of sources and Internet sites
<b>WHY</b>	Historical interest

What happened on the *Titanic* at 11:40 on the night of April 14, 1912, is well known. Frederick Fleet’s cry of “Iceberg, right ahead” and the three accompanying pulls of the crow’s nest bell signaled the beginning of a nightmare that has become legend. By 2:15 a.m., the *Titanic*, thought by many to be unsinkable, had sunk, leaving more than 1500 passengers and crew members on board to meet their icy fate.

Here are some data about the passengers and crew aboard the *Titanic*. Each case (row) of the data table represents a person on board the ship. The variables are the person’s *Survival* status (Dead or Alive), *Age* (Adult or Child), *Sex* (Male or Female), and ticket *Class* (First, Second, Third, or Crew).

The problem with a data table like this—and in fact with all data tables—is that you can’t see what’s going on. And seeing is just what we want to do. We need ways to show the data so that we can see patterns, relationships, trends, and exceptions.

**AS** **Video: The Incident** tells the story of the *Titanic*, and includes rare film footage.

Survival	Age	Sex	Class
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Third
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Dead	Adult	Male	Crew
Alive	Adult	Female	First
Dead	Adult	Male	Third
Dead	Adult	Male	Crew

**Table 3.1**

Part of a data table showing four variables for nine people aboard the *Titanic*.

# The Three Rules of Data Analysis



**FIGURE 3.1 A Picture to Tell a Story**

Florence Nightingale (1820–1910), a founder of modern nursing, was also a pioneer in health management, statistics, and epidemiology. She was the first female member of the British Statistical Society and was granted honorary membership in the newly formed American Statistical Association.

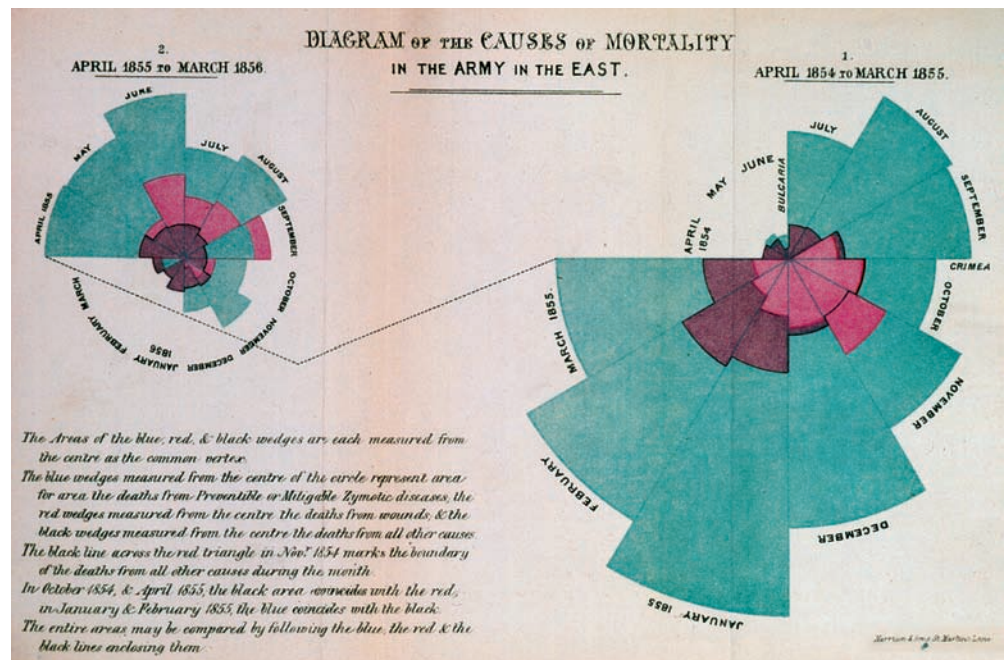
To argue forcefully for better hospital conditions for soldiers, she and her colleague, Dr. William Farr, invented this display, which showed that in the Crimean War, far more soldiers died of illness and infection than of battle wounds. Her campaign succeeded in improving hospital conditions and nursing for soldiers.

Florence Nightingale went on to apply statistical methods to a variety of important health issues and published more than 200 books, reports, and pamphlets during her long and illustrious career.

So, what should we do with data like these? There are three things you should always do first with data:

1. **Make a picture.** A display of your data will reveal things you are not likely to see in a table of numbers and will help you to *Think* clearly about the patterns and relationships that may be hiding in your data.
2. **Make a picture.** A well-designed display will *Show* the important features and patterns in your data. A picture will also show you the things you did not expect to see: the extraordinary (possibly wrong) data values or unexpected patterns.
3. **Make a picture.** The best way to *Tell* others about your data is with a well-chosen picture.

These are the three rules of data analysis. There are pictures of data throughout the book, and new kinds keep showing up. These days, technology makes drawing pictures of data easy, so there is no reason not to follow the three rules.



# Frequency Tables: Making Piles

**AS** **Activity:** Make and examine a table of counts. Even data on something as simple as hair color can reveal surprises when you organize it in a data table.

Class	Count
First	325
Second	285
Third	706
Crew	885

**Table 3.2**  
A frequency table of the *Titanic* passengers.

To make a picture of data, the first thing we have to do is to make piles. Making piles is the beginning of understanding about data. We pile together things that seem to go together, so we can see how the cases distribute across different categories. For categorical data, piling is easy. We just count the number of cases corresponding to each category and pile them up.

One way to put all 2201 people on the *Titanic* into piles is by ticket *Class*, counting up how many had each kind of ticket. We can organize these counts into a **frequency table**, which records the totals and the category names.

Even when we have thousands of cases, a variable like ticket *Class*, with only a few categories, has a frequency table that's easy to read. A frequency table with dozens or hundreds of categories would be much harder to read. We use the names of the categories to label each row in the frequency table. For ticket *Class*, these are "First," "Second," "Third," and "Crew."



Class	%
First	14.77
Second	12.95
Third	32.08
Crew	40.21

Table 3.3

A relative frequency table for the same data.

Counts are useful, but sometimes we want to know the fraction or **proportion** of the data in each category, so we divide the counts by the total number of cases. Usually we multiply by 100 to express these proportions as **percentages**. A **relative frequency table** displays the *percentages*, rather than the counts, of the values in each category. Both types of tables show how the cases are distributed across the categories. In this way, they describe the **distribution** of a categorical variable because they name the possible categories and tell how frequently each occurs.

## The Area Principle

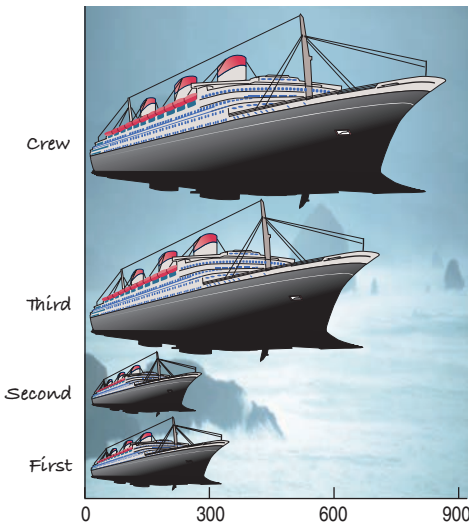


FIGURE 3.2

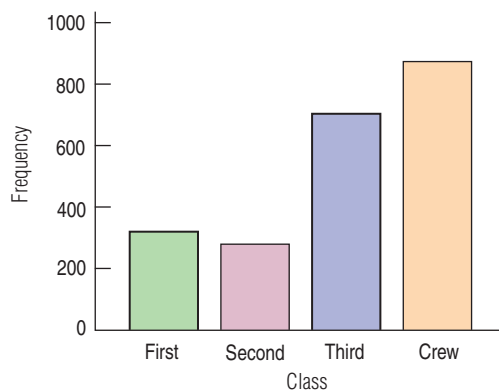
How many people were in each class on the *Titanic*? From this display, it looks as though the service must have been great, since most aboard were crew members. Although the length of each ship here corresponds to the correct number, the impression is all wrong. In fact, only about 40% were crew.

Now that we have the frequency table, we're ready to follow the three rules of data analysis and make a picture of the data. But a bad picture can distort our understanding rather than help it. Here's a graph of the *Titanic* data. What impression do you get about who was aboard the ship?

It sure looks like most of the people on the *Titanic* were crew members, with a few passengers along for the ride. That doesn't seem right. What's wrong? The lengths of the ships *do* match the totals in the table. (You can check the scale at the bottom.) However, experience and psychological tests show that our eyes tend to be more impressed by the *area* than by other aspects of each ship image. So, even though the *length* of each ship matches up with one of the totals, it's the associated *area* in the image that we notice. Since there were about 3 times as many crew as second-class passengers, the ship depicting the number of crew is about 3 times longer than the ship depicting second-class passengers, but it occupies about 9 times the area. As you can see from the frequency table (Table 3.2), that just isn't a correct impression.

The best data displays observe a fundamental principle of graphing data called the **area principle**. The area principle says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Violations of the area principle are a common way to lie (or, since most mistakes are unintentional, we should say err) with Statistics.

## Bar Charts

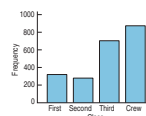
FIGURE 3.3 People on the *Titanic* by Ticket Class

With the area principle satisfied, we can see the true distribution more clearly.

Here's a chart that obeys the area principle. It's not as visually entertaining as the ships, but it does give an *accurate* visual impression of the distribution. The height of each bar shows the count for its category. The bars are the same width, so their heights determine their areas, and the areas are proportional to the counts in each class. Now it's easy to see that the majority of people on board were *not* crew, as the ships picture led us to believe. We can also see that there were about 3 times as many crew as second-class passengers. And there were more than twice as many third-class passengers as either first- or second-class passengers, something you may have missed in the frequency table. Bar charts make these kinds of comparisons easy and natural.

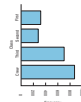
A **bar chart** displays the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison. Bar charts should have small spaces between the bars to indicate that these are freestanding bars that could be rearranged into any order. The bars are lined up along a common base.

Usually they stick up like this



but sometimes they run

sideways like this



If we really want to draw attention to the relative *proportion* of passengers falling into each of these classes, we could replace the counts with percentages and use a **relative frequency bar chart**.

### AS Activity: Bar Charts.

Watch bar charts grow from data; then use your statistics package to create some bar charts for yourself.

For some reason, some computer programs give the name “bar chart” to any graph that uses bars. And others use different names according to whether the bars are horizontal or vertical. Don’t be misled. “Bar chart” is the term for a *display of counts of a categorical variable with bars*.

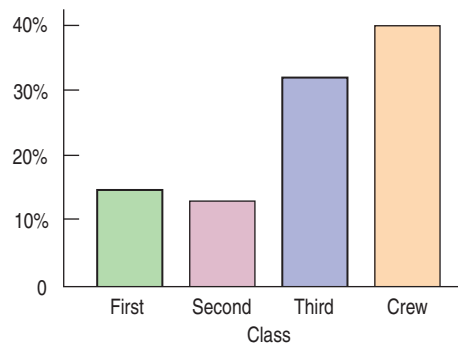


FIGURE 3.4

The relative frequency bar chart looks the same as the bar chart (Figure 3.3) but shows the proportion of people in each category rather than the counts.

## Pie Charts

Another common display that shows how a whole group breaks into several categories is a pie chart. **Pie charts** show the whole group of cases as a circle. They slice the circle into pieces whose sizes are proportional to the fraction of the whole in each category.

Pie charts give a quick impression of how a whole group is partitioned into smaller groups. Because we’re used to cutting up pies into 2, 4, or 8 pieces, pie charts are good for seeing relative frequencies near  $1/2$ ,  $1/4$ , or  $1/8$ . For example, you may be able to tell that the pink slice, representing the second-class passengers, is very close to  $1/8$  of the total. It’s harder to see that there were about twice as many third-class as first-class passengers. Which category had the most passengers? Were there more crew or more third-class passengers? Comparisons such as these are easier in a bar chart.

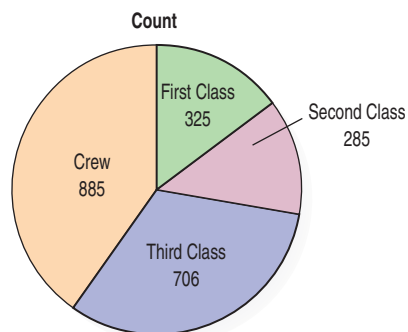


FIGURE 3.5 Number of Titanic passengers in each class

**Think before you draw.** Our first rule of data analysis is *Make a picture*. But what kind of picture? We don’t have a lot of options—yet. There’s more to Statistics than pie charts and bar charts, and knowing when to use each type of graph is a critical first step in data analysis. That decision depends in part on what type of data we have.

It’s important to check that the data are appropriate for whatever method of analysis you choose. **Before you make a bar chart or a pie chart, always check the**

**Categorical Data Condition:** The data are counts or percentages of individuals in categories.

If you want to make a relative frequency bar chart or a pie chart, you'll need to also make sure that the categories don't overlap so that no individual is counted twice. If the categories do overlap, you can still make a bar chart, but the percentages won't add up to 100%. For the *Titanic* data, either kind of display is appropriate because the categories don't overlap.

Throughout this course, you'll see that doing Statistics right means selecting the proper methods. That means you have to *Think* about the situation at hand. An important first step, then, is to check that the type of analysis you plan is appropriate. The Categorical Data Condition is just the first of many such checks.

## Contingency Tables: Children and First-Class Ticket Holders First?

**AS** **Activity: Children at Risk.**  
This activity looks at the fates of children aboard the *Titanic*; the subsequent activity shows how to make such tables on a computer.

We know how many tickets of each class were sold on the *Titanic*, and we know that only about 32% of all those aboard the *Titanic* survived. After looking at the distribution of each variable by itself, it's natural and more interesting to ask how they relate. Was there a relationship between the kind of ticket a passenger held and the passenger's chances of making it into the lifeboat? To answer this question, we need to look at the two categorical variables *Class* and *Survival* together.

To look at two categorical variables together, we often arrange the counts in a two-way table. Here is a two-way table of those aboard the *Titanic*, classified according to the class of ticket and whether the ticket holder survived or didn't. Because the table shows how the individuals are distributed along each variable, contingent on the value of the other variable, such a table is called a **contingency table**.

Contingency table of ticket *Class* and *Survival*. The bottom line of "Totals" is the same as the previous frequency table.

Table 3.4

		Class				Total
		First	Second	Third	Crew	
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201

The margins of the table, both on the right and at the bottom, give totals. The bottom line of the table is just the frequency distribution of ticket *Class*. The right column of the table is the frequency distribution of the variable *Survival*. When presented like this, in the margins of a contingency table, the frequency distribution of one of the variables is called its **marginal distribution**.

Each **cell** of the table gives the count for a combination of values of the two variables. If you look down the column for second-class passengers to the first cell, you can see that 118 second-class passengers survived. Looking at the third-class passengers, you can see that more third-class passengers (178) survived. Were second-class passengers more likely to survive? Questions like this are easier to address by using percentages. The 118 survivors in second class were 41.4% of the total 285 second-class passengers, while the 178 surviving third-class passengers were only 25.2% of that class's total.

We know that 118 second-class passengers survived. We could display this number as a percentage—but as a percentage of what? The total number of passengers? (118 is 5.4% of the total: 2201.) The number of second-class passengers?



A bell-shaped artifact from the *Titanic*.

(118 is 41.4% of the 285 second-class passengers.) The number of survivors? (118 is 16.6% of the 711 survivors.) All of these are possibilities, and all are potentially useful or interesting. You'll probably wind up calculating (or letting your technology calculate) lots of percentages. Most statistics programs offer a choice of total percent, row percent, or column percent for contingency tables. Unfortunately, they often put them all together with several numbers in each cell of the table. The resulting table holds lots of information, but it can be hard to understand:

Another contingency table of ticket Class. This time we see not only the counts for each combination of Class and Survival (in bold) but the percentages these counts represent. For each count, there are three choices for the percentage: by row, by column, and by table total. There's probably too much information here for this table to be useful.

**Table 3.5**

		Class					
		First	Second	Third	Crew	Total	
Survival	Alive	Count	<b>203</b>	<b>118</b>	<b>178</b>	<b>212</b>	<b>711</b>
		% of Row	28.6%	16.6%	25.0%	29.8%	100%
		% of Column	62.5%	41.4%	25.2%	24.0%	32.3%
		% of Table	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	Count	<b>122</b>	<b>167</b>	<b>528</b>	<b>673</b>	<b>1490</b>
		% of Row	8.2%	11.2%	35.4%	45.2%	100%
		% of Column	37.5%	58.6%	74.8%	76.0%	67.7%
		% of Table	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	Count	<b>325</b>	<b>285</b>	<b>706</b>	<b>885</b>	<b>2201</b>
		% of Row	14.8%	12.9%	32.1%	40.2%	100%
		% of Column	100%	100%	100%	100%	100%
		% of Table	14.8%	12.9%	32.1%	40.2%	100%

To simplify the table, let's first pull out the percent of table values:

A contingency table of Class by Survival with only the table percentages

**Table 3.6**

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	9.2%	5.4%	8.1%	9.6%	32.3%
	Dead	5.6%	7.6%	24.0%	30.6%	67.7%
	Total	14.8%	12.9%	32.1%	40.2%	100%

These percentages tell us what percent of *all* passengers belong to each combination of column and row category. For example, we see that although 8.1% of the people aboard the *Titanic* were surviving third-class ticket holders, only 5.4% were surviving second-class ticket holders. Is this fact useful? Comparing these percentages, you might think that the chances of surviving were better in third class than in second. But be careful. There were many more third-class than second-class passengers on the *Titanic*, so there were more third-class survivors. That group is a larger percentage of the passengers, but is that really what we want to know?

**Percent of what?** The English language can be tricky when we talk about percentages. If you're asked "What percent of the survivors were in second class?" it's pretty clear that we're interested only in survivors. It's as if we're restricting the *Who* in the question to the survivors, so we should look at the number of second-class passengers among all the survivors—in other words, the row percent.

But if you're asked "What percent were second-class passengers who survived?" you have a different question. Be careful; here, the *Who* is everyone on board, so 2201 should be the denominator, and the answer is the table percent.

And if you're asked "What percent of the second-class passengers survived?" you have a third question. Now the *Who* is the second-class passengers, so the denominator is the 285 second-class passengers, and the answer is the column percent.

Always be sure to ask "percent of what?" That will help you to know the *Who* and whether we want *row*, *column*, or *table* percentages.

## FOR EXAMPLE

### Finding marginal distributions

In January 2007, a Gallup poll asked 1008 Americans age 18 and over whether they planned to watch the upcoming Super Bowl. The pollster also asked those who planned to watch whether they were looking forward more to seeing the football game or the commercials. The results are summarized in the table:

**Question:** What's the marginal distribution of the responses?

To determine the percentages for the three responses, divide the count for each response by the total number of people polled:

$$\frac{479}{1008} = 47.5\% \quad \frac{237}{1008} = 23.5\% \quad \frac{292}{1008} = 29.0\%$$

According to the poll, 47.5% of American adults were looking forward to watching the Super Bowl game, 23.5% were looking forward to watching the commercials, and 29% didn't plan to watch at all.

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
Total		492	516	1008

## Conditional Distributions

The more interesting questions are *contingent*. We'd like to know, for example, what percentage of *second-class passengers* survived and how that compares with the survival rate for third-class passengers.

It's more interesting to ask whether the chance of surviving the *Titanic* sinking *depended* on ticket class. We can look at this question in two ways. First, we could ask how the distribution of ticket *Class* changes between survivors and non-survivors. To do that, we look at the *row percentages*:

The conditional distribution of ticket *Class* conditioned on each value of *Survival*: *Alive* and *Dead*.

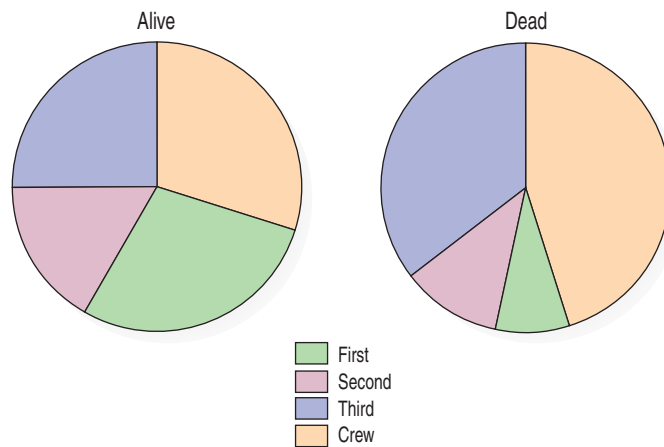
Table 3.7

		Class				Total
		First	Second	Third	Crew	
Survival	Alive	203 28.6%	118 16.6%	178 25.0%	212 29.8%	711 100%
	Dead	122 8.2%	167 11.2%	528 35.4%	673 45.2%	1490 100%

By focusing on each row separately, we see the distribution of class under the *condition* of surviving or not. The sum of the percentages in each row is 100%, and we divide that up by ticket class. In effect, we temporarily restrict the *Who* first to survivors and make a pie chart for them. Then we refocus the *Who* on the nonsurvivors and make their pie chart. These pie charts show the distribution of ticket classes *for each row* of the table: survivors and nonsurvivors. The distributions we create this way are called **conditional distributions**, because they show the distribution of one variable for just those cases that satisfy a condition on another variable.

**FIGURE 3.6**

Pie charts of the conditional distributions of ticket Class for the survivors and nonsurvivors, separately. Do the distributions appear to be the same? We're primarily concerned with percentages here, so pie charts are a reasonable choice.



**FOR EXAMPLE** Finding conditional distributions

**Recap:** The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

**Question:** How do the conditional distributions of interest in the commercials differ for men and women?

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
	Total	492	516	1008

Look at the group of people who responded "Commercials" and determine what percent of them were male and female:

$$\frac{81}{237} = 34.2\% \quad \frac{156}{237} = 65.8\%$$

Women make up a sizable majority of the adult Americans who look forward to seeing Super Bowl commercials more than the game itself. Nearly 66% of people who voiced a preference for the commercials were women, and only 34% were men.

But we can also turn the question around. We can look at the distribution of *Survival* for each category of ticket *Class*. To do this, we look at the *column percentages*. Those show us whether the chance of surviving was roughly the same for each of the four classes. Now the percentages in each column add to 100%, because we've restricted the *Who*, in turn, to each of the four ticket classes:

A contingency table of *Class* by *Survival* with only counts and column percentages. Each column represents the conditional distribution of *Survival* for a given category of ticket *Class*.

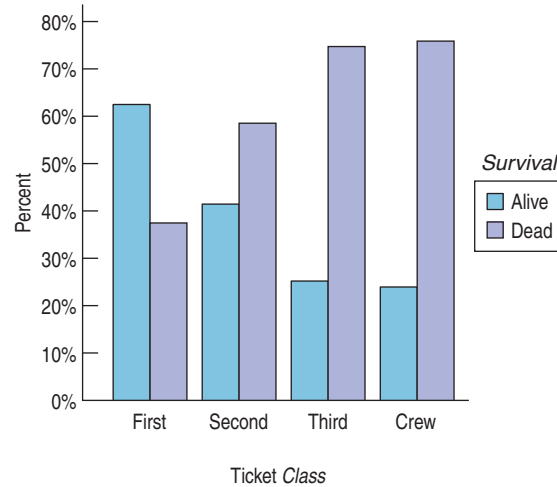
**Table 3.8**

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	Count 203 % of Column 62.5%	Count 118 % of Column 41.4%	Count 178 % of Column 25.2%	Count 212 % of Column 24.0%	Count 711 % of Column 32.3%
	Dead	Count 122 % of Column 37.5%	Count 167 % of Column 58.6%	Count 528 % of Column 74.8%	Count 673 % of Column 76.0%	Count 1490 % of Column 67.7%
	Total	Count 325 100%	Count 285 100%	Count 706 100%	Count 885 100%	Count 2201 100%

Looking at how the percentages change across each row, it sure looks like ticket class mattered in whether a passenger survived. To make it more vivid, we could show the distribution of *Survival* for each ticket class in a display. Here's a side-by-side bar chart showing percentages of surviving and not for each category:

**FIGURE 3.7**

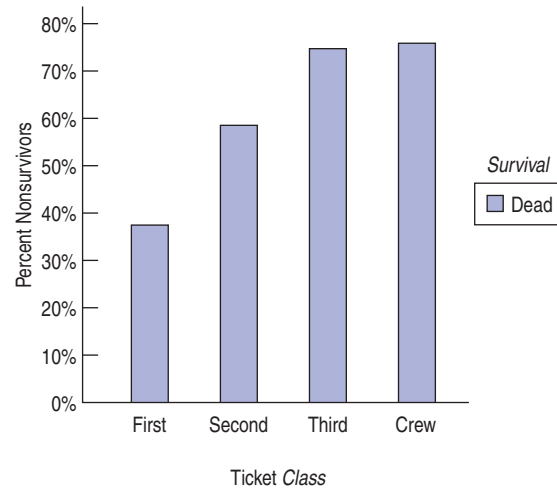
**Side-by-side bar chart** showing the conditional distribution of *Survival* for each category of ticket *Class*. The corresponding pie charts would have only two categories in each of four pies, so bar charts seem the better alternative.



These bar charts are simple because, for the variable *Survival*, we have only two alternatives: Alive and Dead. When we have only two categories, we really need to know only the percentage of one of them. Knowing the percentage that survived tells us the percentage that died. We can use this fact to simplify the display even more by dropping one category. Here are the percentages of dying across the classes displayed in one chart:

**FIGURE 3.8**

**Bar chart** showing just nonsurvivor percentages for each value of ticket *Class*. Because we have only two values, the second bar doesn't add any information. Compare this chart to the side-by-side bar chart shown earlier.



### TI-*nspire*

**Conditional distributions and association.** Explore the *Titanic* data to see which passengers were most likely to survive.

Now it's easy to compare the risks. Among first-class passengers, 37.5% perished, compared to 58.6% for second-class ticket holders, 74.8% for those in third class, and 76.0% for crew members.

If the risk had been about the same across the ticket classes, we would have said that survival was *independent* of class. But it's not. The differences we see among these conditional distributions suggest that survival may have depended on ticket class. You may find it useful to consider conditioning on each variable in a contingency table in order to explore the dependence between them.

It is interesting to know that *Class* and *Survival* are associated. That's an important part of the *Titanic* story. And we know how important this is because the margins show us the actual numbers of people involved.

Variables can be associated in many ways and to different degrees. The best way to tell whether two variables are associated is to ask whether they are *not*.<sup>1</sup> In a contingency table, when the distribution of *one* variable is the same for all categories of another, we say that the variables are **independent**. That tells us there's no association between these variables. We'll see a way to check for independence formally later in the book. For now, we'll just compare the distributions.

## FOR EXAMPLE

### Looking for associations between variables

**Recap:** The table shows results of a poll asking adults whether they were looking forward to the Super Bowl game, looking forward to the commercials, or didn't plan to watch.

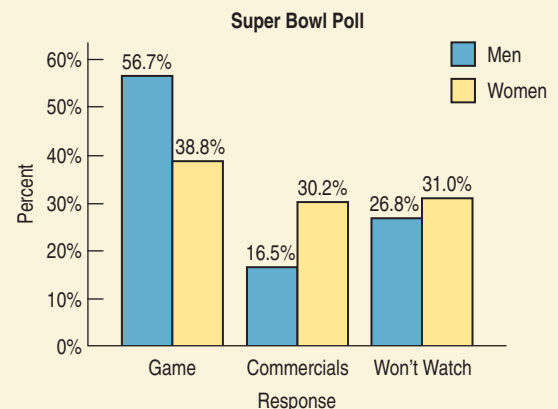
**Question:** Does it seem that there's an association between interest in Super Bowl TV coverage and a person's sex?

		Sex		
		Male	Female	Total
Response	Game	279	200	479
	Commercials	81	156	237
	Won't watch	132	160	292
	Total	492	516	1008

First find the distribution of the three responses for the men (the column percentages):

$$\frac{279}{492} = 56.7\% \quad \frac{81}{492} = 16.5\% \quad \frac{132}{492} = 26.8\%$$

Then do the same for the women who were polled, and display the two distributions with a side-by-side bar chart:



Based on this poll it appears that women were only slightly less interested than men in watching the Super Bowl telecast: 31% of the women said they didn't plan to watch, compared to just under 27% of men. Among those who planned to watch, however, there appears to be an association between the viewer's sex and what the viewer is most looking forward to. While more women are interested in the game (39%) than the commercials (30%), the margin among men is much wider: 57% of men said they were looking forward to seeing the game, compared to only 16.5% who cited the commercials.

<sup>1</sup>This kind of "backwards" reasoning shows up surprisingly often in science—and in Statistics. We'll see it again.





## JUST CHECKING

A Statistics class reports the following data on Sex and Eye Color for students in the class:

		Eye Color			Total
		Blue	Brown	Green/Hazel/Other	
Sex	Males	6	20	6	32
	Females	4	16	12	32
	Total	10	36	18	64

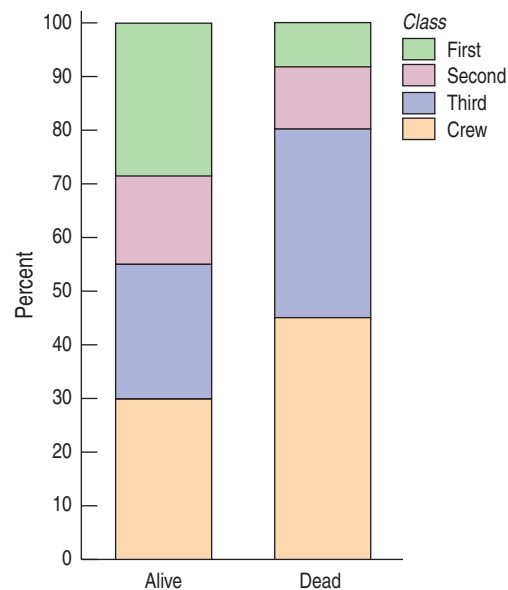
1. What percent of females are brown-eyed?
2. What percent of brown-eyed students are female?
3. What percent of students are brown-eyed females?
4. What's the distribution of Eye Color?
5. What's the conditional distribution of Eye Color for the males?
6. Compare the percent who are female among the blue-eyed students to the percent of all students who are female.
7. Does it seem that Eye Color and Sex are independent? Explain.

## Segmented Bar Charts

We could display the *Titanic* information by dividing up bars rather than circles. The resulting **segmented bar chart** treats each bar as the “whole” and divides it proportionally into segments corresponding to the percentage in each group. We can clearly see that the distributions of ticket *Class* are different, indicating again that survival was not independent of ticket *Class*.

**FIGURE 3.9** A segmented bar chart for Class by Survival

Notice that although the totals for survivors and nonsurvivors are quite different, the bars are the same height because we have converted the numbers to percentages. Compare this display with the side-by-side pie charts of the same data in Figure 3.6.



## STEP-BY-STEP EXAMPLE

## Examining Contingency Tables

Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer (“Fatty Fish Consumption and Risk of Prostate Cancer,” *Lancet*, June 2001). Their results are summarized in this table:



We asked for a picture of a man eating fish. This is what we got.

		Prostate Cancer	
		No	Yes
Fish Consumption	Never/seldom	110	14
	Small part of diet	2420	201
	Moderate part	2769	209
	Large part	507	42

Table 3.9

**Question:** Is there an association between fish consumption and prostate cancer?



**Plan** Be sure to state what the problem is about.

**Variables** Identify the variables and report the W's.

Be sure to check the appropriate condition.

I want to know if there is an association between fish consumption and prostate cancer.

The individuals are 6272 Swedish men followed by medical researchers for 30 years. The variables record their fish consumption and whether or not they were diagnosed with prostate cancer.

✓ **Categorical Data Condition:** I have counts for both fish consumption and cancer diagnosis. The categories of diet do not overlap, and the diagnoses do not overlap. It's okay to draw pie charts or bar charts.

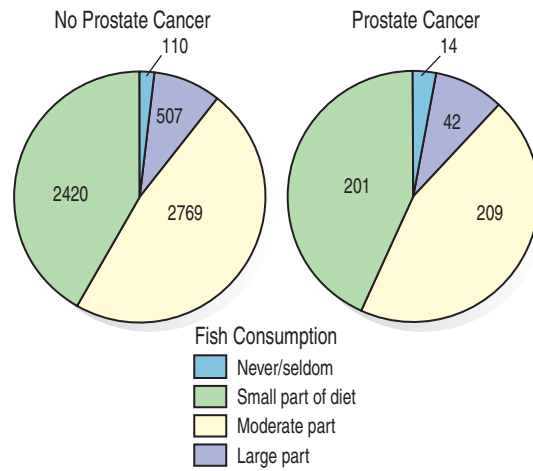


**Mechanics** It's a good idea to check the marginal distributions first before looking at the two variables together.

		Prostate Cancer		
		No	Yes	Total
Fish Consumption	Never/seldom	110	14	124 (2.0%)
	Small part of diet	2420	201	2621 (41.8%)
	Moderate part	2769	209	2978 (47.5%)
	Large part	507	42	549 (8.8%)
	Total	5806 (92.6%)	466 (7.4%)	6272 (100%)

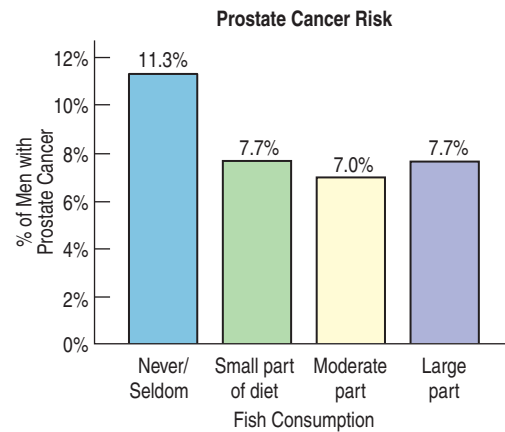
Two categories of the diet are quite small, with only 2.0% Never/Seldom eating fish and 8.8% in the “Large part” category. Overall, 7.4% of the men in this study had prostate cancer.

Then, make appropriate displays to see whether there is a difference in the relative proportions. These pie charts compare fish consumption for men who have prostate cancer to fish consumption for men who don't.



It's hard to see much difference in the pie charts. So, I made a display of the row percentages. Because there are only two alternatives, I chose to display the risk of prostate cancer for each group:

Both pie charts and bar charts can be used to compare conditional distributions. Here we compare prostate cancer rates based on differences in fish consumption.



**Conclusion** Interpret the patterns in the table and displays in context. If you can, discuss possible real-world consequences. Be careful not to overstate what you see. The results may not generalize to other situations.

Overall, there is a 7.4% rate of prostate cancer among men in this study. Most of the men (89.3%) ate fish either as a moderate or small part of their diet. From the pie charts, it's hard to see a difference in cancer rates among the groups. But in the bar chart, it looks like the cancer rate for those who never/seldom ate fish may be somewhat higher.

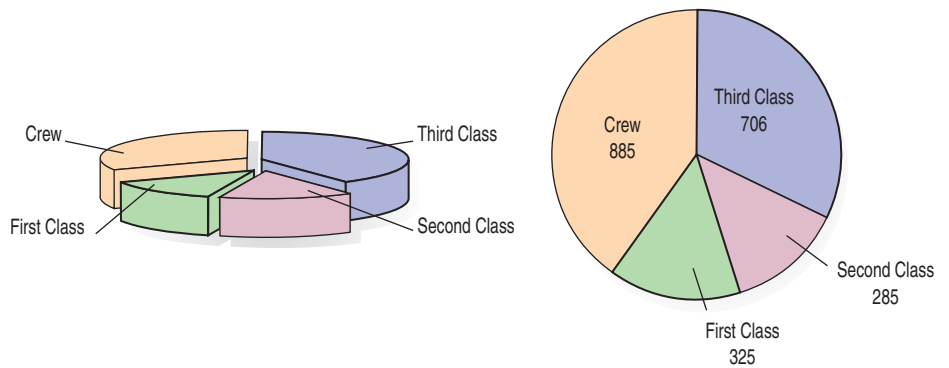
However, only 124 of the 6272 men in the study fell into this category, and only 14 of them developed prostate cancer. More study would probably be needed before we would recommend that men change their diets.<sup>2</sup>

<sup>2</sup> The original study actually used pairs of twins, which enabled the researchers to discern that the risk of cancer for those who never ate fish actually *was* substantially greater. Using pairs is a special way of gathering data. We'll discuss such study design issues and how to analyze the data in the later chapters.

This study is an example of looking at a sample of data to learn something about a larger population. We care about more than these particular 6272 Swedish men. We hope that learning about their experiences will tell us something about the value of eating fish in general. That raises the interesting question of what population we think this sample might represent. Do we hope to learn about all Swedish men? About all men? About the value of eating fish for all adult humans? <sup>3</sup> Often, it can be hard to decide just which population our findings may tell us about, but that also is how researchers decide what to look into in future studies.

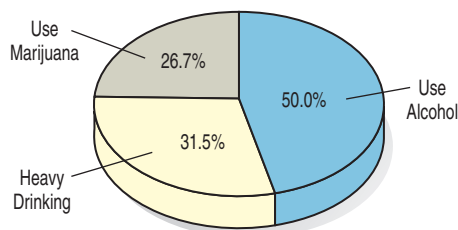
### WHAT CAN GO WRONG?

- ▶ **Don't violate the area principle.** This is probably the most common mistake in a graphical display. It is often made in the cause of artistic presentation. Here, for example, are two displays of the pie chart of the *Titanic* passengers by class:



The one on the left looks pretty, doesn't it? But showing the pie on a slant violates the area principle and makes it much more difficult to compare fractions of the whole made up of each class—the principal feature that a pie chart ought to show.

- ▶ **Keep it honest.** Here's a pie chart that displays data on the percentage of high school students who engage in specified dangerous behaviors as reported by the Centers for Disease Control. What's wrong with this plot?

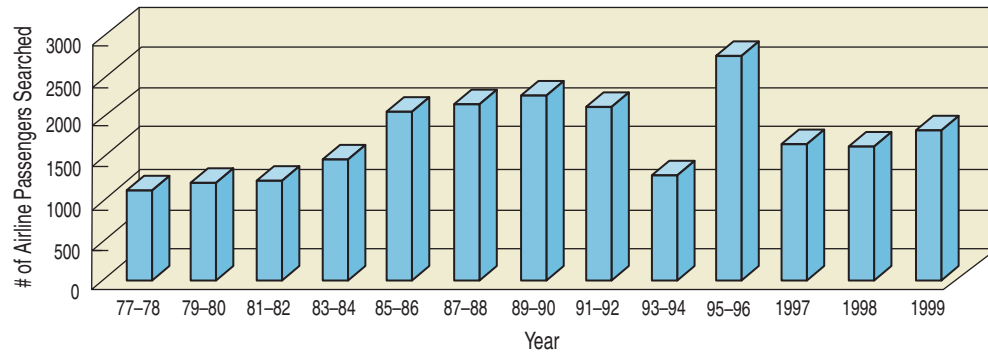


Try adding up the percentages. Or look at the 50% slice. Does it look right? Then think: What are these percentages of? Is there a "whole" that has been sliced up? In a pie chart, the proportions shown by each slice of the pie must add up to 100% and each individual must fall into only one category. Of course, showing the pie on a slant makes it even harder to detect the error.

(continued)

<sup>3</sup> Probably not, since we're looking only at prostate cancer risk.

Here's another. This bar chart shows the number of airline passengers searched in security screening, by year:



Looks like things didn't change much in the final years of the 20th century—until you read the bar labels and see that the last three bars represent single years while all the others are for *pairs* of years. Of course, the false depth makes it harder to see the problem.

- ▶ **Don't confuse similar-sounding percentages.** These percentages sound similar but are different:
  - ▶ The percentage of the passengers who were both in first class and survived: This would be 203/2201, or 9.4%.
  - ▶ The percentage of the first-class passengers who survived: This is 203/325, or 62.5%.
  - ▶ The percentage of the survivors who were in first class: This is 203/711, or 28.6%.

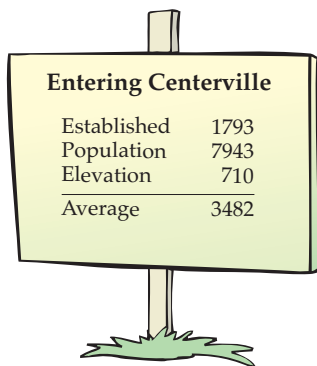
In each instance, pay attention to the *Who* implicitly defined by the phrase. Often there is a restriction to a smaller group (all aboard the *Titanic*, those in first class, and those who survived, respectively) before a percentage is found. Your discussion of results must make these differences clear.

- ▶ **Don't forget to look at the variables separately, too.** When you make a contingency table or display a conditional distribution, be sure you also examine the marginal distributions. It's important to know how many cases are in each category.
- ▶ **Be sure to use enough individuals.** When you consider percentages, take care that they are based on a large enough number of individuals. Take care not to make a report such as this one:

*We found that 66.67% of the rats improved their performance with training. The other rat died.*

- ▶ **Don't overstate your case.** Independence is an important concept, but it is rare for two variables to be *entirely* independent. We can't conclude that one variable has no effect whatsoever on another. Usually, all we know is that little effect was observed in our study. Other studies of other groups under other circumstances could find different results.

		Class				
		First	Second	Third	Crew	Total
Survival	Alive	203	118	178	212	711
	Dead	122	167	528	673	1490
	Total	325	285	706	885	2201



### SIMPSON'S PARADOX

- ▶ **Don't use unfair or silly averages.** Sometimes averages can be misleading. Sometimes they just don't make sense at all. Be careful when averaging different variables that the quantities you're averaging are comparable. The Centerville sign says it all.

When using averages of proportions across several different groups, it's important to make sure that the groups really are comparable.

It's easy to make up an example showing that averaging across very different values or groups can give absurd results. Here's how that might work: Suppose there are two pilots, Moe and Jill. Moe argues that he's the better pilot of the two, since he managed to land 83% of his last 120 flights on time compared with Jill's 78%. But let's look at the data a little more closely. Here are the results for each of their last 120 flights, broken down by the time of day they flew:

**Table 3.10**

On-time flights by *Time of Day* and *Pilot*. Look at the percentages within each *Time of Day* category. Who has a better on-time record during the day? At night? Who is better overall?

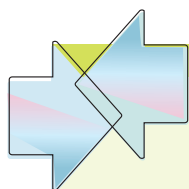
		Time of Day		
		Day	Night	Overall
Pilot	Moe	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Jill	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

One famous example of Simpson's paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% of female applicants got in. It looked like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc.), it turned out that, within each school, the women were admitted at nearly the same or, in some cases, much *higher* rates than the men. How could this be? Women applied in large numbers to schools with very low admission rates (Law and Medicine, for example, admitted fewer than 10%). Men tended to apply to Engineering and Science. Those schools have admission rates above 50%. When the *average* was taken, the women had a much lower *overall* rate, but the average didn't really make sense.

Look at the daytime and nighttime flights separately. For day flights, Jill had a 95% on-time rate and Moe only a 90% rate. At night, Jill was on time 75% of the time and Moe only 50%. So Moe is better "overall," but Jill is better both during the day and at night. How can this be?

What's going on here is a problem known as **Simpson's paradox**, named for the statistician who discovered it in the 1960s. It comes up rarely in real life, but there have been several well-publicized cases. As we can see from the pilot example, the problem is *unfair averaging* over different groups. Jill has mostly night flights, which are more difficult, so her *overall average* is heavily influenced by her nighttime average. Moe, on the other hand, benefits from flying mostly during the day, with its higher on-time percentage. With their very different patterns of flying conditions, taking an overall average is misleading. It's not a fair comparison.

The moral of Simpson's paradox is to be careful when you average across different levels of a second variable. It's always better to compare percentages or other averages *within* each level of the other variable. The overall average may be misleading.



## CONNECTIONS

All of the methods of this chapter work with *categorical variables*. You must know the *Who* of the data to know who is counted in each category and the *What* of the variable to know where the categories come from.



## WHAT HAVE WE LEARNED?

We've learned that we can summarize categorical data by counting the number of cases in each category, sometimes expressing the resulting distribution as percents. We can display the distribution in a bar chart or a pie chart. When we want to see how two categorical variables are related, we put the counts (and/or percentages) in a two-way table called a contingency table.

- ▶ We look at the marginal distribution of each variable (found in the margins of the table).
- ▶ We also look at the conditional distribution of a variable within each category of the other variable.
- ▶ We can display these conditional and marginal distributions by using bar charts or pie charts.
- ▶ If the conditional distributions of one variable are (roughly) the same for every category of the other, the variables are independent.

### Terms

Frequency table  
(Relative frequency table)

Distribution

Area principle

Bar chart  
(Relative frequency bar chart)

Pie chart

Categorical data condition

Contingency table

Marginal distribution

Conditional distribution

Independence

Segmented bar chart

Simpson's paradox

21. A frequency table lists the categories in a categorical variable and gives the count (or percentage) of observations for each category.

22. The distribution of a variable gives

- ▶ the possible values of the variable and
- ▶ the relative frequency of each value.

22. In a statistical display, each data value should be represented by the same amount of area.

22. Bar charts show a bar whose area represents the count (or percentage) of observations for each category of a categorical variable.

23. Pie charts show how a "whole" divides into categories by showing a wedge of a circle whose area corresponds to the proportion in each category.

24. The methods in this chapter are appropriate for displaying and describing categorical data. Be careful not to use them with quantitative data.

24. A contingency table displays counts and, sometimes, percentages of individuals falling into named categories on two or more variables. The table categorizes the individuals on all variables at once to reveal possible patterns in one variable that may be contingent on the category of the other.

24. In a contingency table, the distribution of either variable alone is called the marginal distribution. The counts or percentages are the totals found in the margins (last row or column) of the table.

26. The distribution of a variable restricting the *Who* to consider only a smaller group of individuals is called a conditional distribution.

29. Variables are said to be independent if the conditional distribution of one variable is the same for each category of the other. We'll show how to check for independence in a later chapter.

30. A segmented bar chart displays the conditional distribution of a categorical variable within each category of another variable.

34. When averages are taken across different groups, they can appear to contradict the overall averages. This is known as "Simpson's paradox."

### Skills

THINK

- ▶ Be able to recognize when a variable is categorical and choose an appropriate display for it.
- ▶ Understand how to examine the association between categorical variables by comparing conditional and marginal percentages.

SHOW

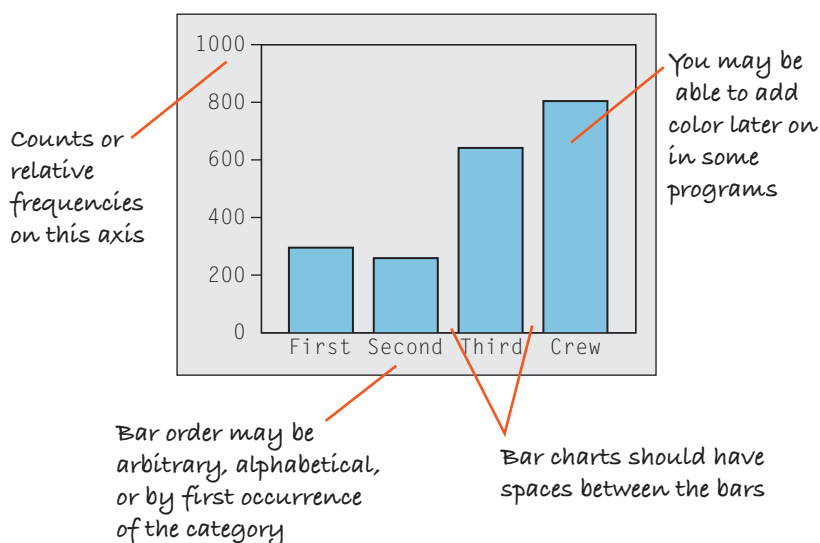
- ▶ Be able to summarize the distribution of a categorical variable with a frequency table.
- ▶ Be able to display the distribution of a categorical variable with a bar chart or pie chart.
- ▶ Know how to make and examine a contingency table.



- ▶ Know how to make and examine displays of the conditional distributions of one variable for two or more groups.
- ▶ Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- ▶ Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

## DISPLAYING CATEGORICAL DATA ON THE COMPUTER

Although every package makes a slightly different bar chart, they all have similar features:



Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

## EXERCISES

1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
  - a) Is the graph clearly labeled?
  - b) Does it violate the area principle?
  - c) Does the accompanying article tell the W's of the variable?
  - d) Do you think the article correctly interprets the data? Explain.
2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
  - a) Is the graph clearly labeled?
  - b) Does it violate the area principle?
  - c) Does the accompanying article tell the W's of the variable?
  - d) Do you think the article correctly interprets the data? Explain.

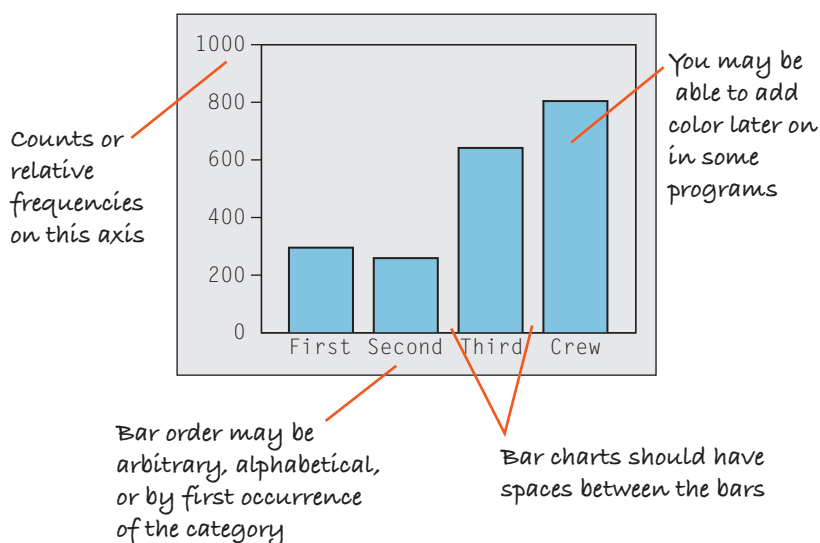




- ▶ Know how to make and examine displays of the conditional distributions of one variable for two or more groups.
- ▶ Be able to describe the distribution of a categorical variable in terms of its possible values and relative frequencies.
- ▶ Know how to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Be able to describe and discuss patterns found in a contingency table and associated displays of conditional distributions.

## DISPLAYING CATEGORICAL DATA ON THE COMPUTER

Although every package makes a slightly different bar chart, they all have similar features:



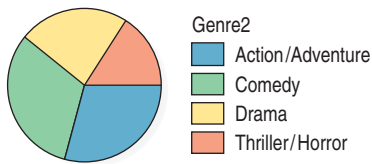
Sometimes the count or a percentage is printed above or on top of each bar to give some additional information. You may find that your statistics package sorts category names in annoying orders by default. For example, many packages sort categories alphabetically or by the order the categories are seen in the data set. Often, neither of these is the best choice.

## EXERCISES

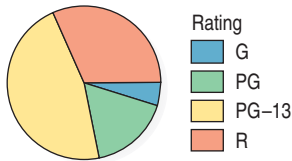
1. **Graphs in the news.** Find a bar graph of categorical data from a newspaper, a magazine, or the Internet.
  - a) Is the graph clearly labeled?
  - b) Does it violate the area principle?
  - c) Does the accompanying article tell the W's of the variable?
  - d) Do you think the article correctly interprets the data? Explain.
2. **Graphs in the news II.** Find a pie chart of categorical data from a newspaper, a magazine, or the Internet.
  - a) Is the graph clearly labeled?
  - b) Does it violate the area principle?
  - c) Does the accompanying article tell the W's of the variable?
  - d) Do you think the article correctly interprets the data? Explain.

3. **Tables in the news.** Find a frequency table of categorical data from a newspaper, a magazine, or the Internet.
- Is it clearly labeled?
  - Does it display percentages or counts?
  - Does the accompanying article tell the *W*'s of the variable?
  - Do you think the article correctly interprets the data? Explain.
4. **Tables in the news II.** Find a contingency table of categorical data from a newspaper, a magazine, or the Internet.
- Is it clearly labeled?
  - Does it display percentages or counts?
  - Does the accompanying article tell the *W*'s of the variables?
  - Do you think the article correctly interprets the data? Explain.

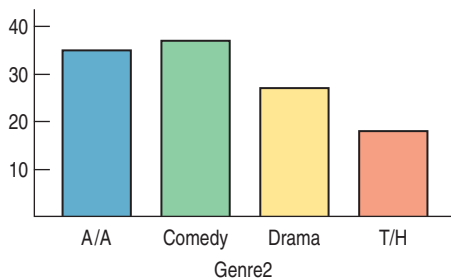
- T** 5. **Movie genres.** The pie chart summarizes the genres of 120 first-run movies released in 2005.
- Is this an appropriate display for the genres? Why/why not?
  - Which genre was least common?



- T** 6. **Movie ratings.** The pie chart shows the ratings assigned to 120 first-run movies released in 2005.
- Is this an appropriate display for these data? Explain.
  - Which was the most common rating?

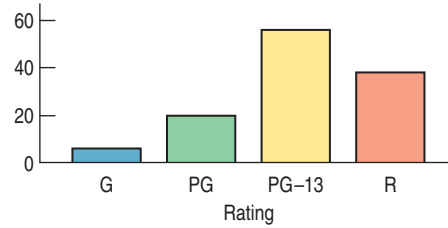


- T** 7. **Genres again.** Here is a bar chart summarizing the 2005 movie genres, as seen in the pie chart in Exercise 5.
- Which genre was most common?
  - Is it easier to see that in the pie chart or the bar chart? Explain.



- T** 8. **Ratings again.** Here is a bar chart summarizing the 2005 movie ratings, as seen in the pie chart in Exercise 6.
- Which was the least common rating?
  - An editorial claimed that there's been a growth in PG-13 rated films that, according to the writer, "have too much sex and violence," at the expense of G-rated

films that offer "good, clean fun." The writer offered the bar chart below as evidence to support his claim. Does the bar chart support his claim? Explain.



9. **Magnet schools.** An article in the Winter 2003 issue of *Chance* magazine reported on the Houston Independent School District's magnet schools programs. Of the 1755 qualified applicants, 931 were accepted, 298 were wait-listed, and 526 were turned away for lack of space. Find the relative frequency distribution of the decisions made, and write a sentence describing it.
10. **Magnet schools again.** The *Chance* article about the Houston magnet schools program described in Exercise 9 also indicated that 517 applicants were black or Hispanic, 292 Asian, and 946 white. Summarize the relative frequency distribution of ethnicity with a sentence or two (in the proper context, of course).
11. **Causes of death 2004.** The Centers for Disease Control and Prevention ([www.cdc.gov](http://www.cdc.gov)) lists causes of death in the United States during 2004:

Cause of Death	Percent
Heart disease	27.2
Cancer	23.1
Circulatory diseases and stroke	6.3
Respiratory diseases	5.1
Accidents	4.7

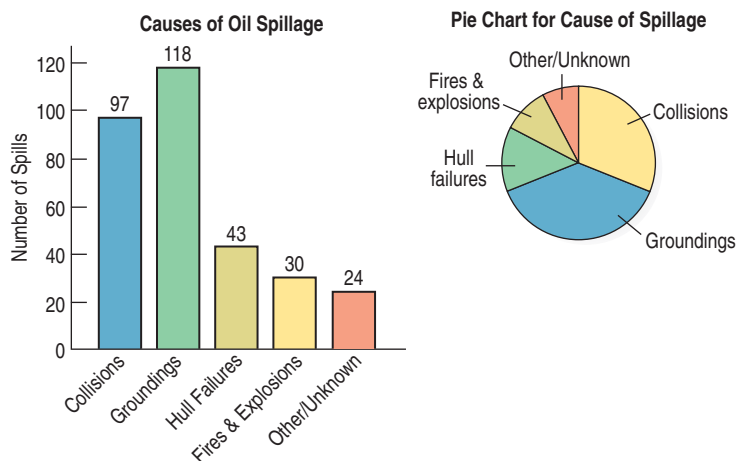
- Is it reasonable to conclude that heart or respiratory diseases were the cause of approximately 33% of U.S. deaths in 2003?
  - What percent of deaths were from causes not listed here?
  - Create an appropriate display for these data.
12. **Plane crashes.** An investigation compiled information about recent nonmilitary plane crashes ([www.planecrashinfo.com](http://www.planecrashinfo.com)). The causes, to the extent that they could be determined, are summarized in the table.

Cause	Percent
Pilot error	40
Other human error	5
Weather	6
Mechanical failure	14
Sabotage	6

- Is it reasonable to conclude that the weather or mechanical failures caused only about 20% of recent plane crashes?
- In what percent of crashes were the causes not determined?
- Create an appropriate display for these data.

13. **Oil spills 2006.** Data from the International Tanker Owners Pollution Federation Limited ([www.itopf.com](http://www.itopf.com)) give the cause of spillage for 312 large oil tanker accidents from 1974–2006. Here are displays.

- Write a brief report interpreting what the displays show.
- Is a pie chart an appropriate display for these data? Why or why not?

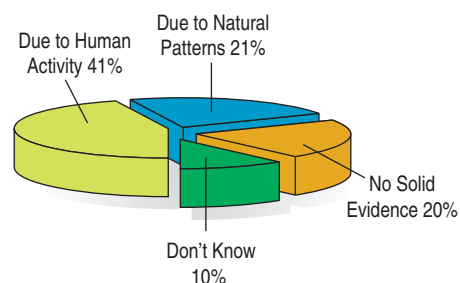


14. **Winter Olympics 2006.** Twenty-six countries won medals in the 2006 Winter Olympics. The table lists them, along with the total number of medals each won:

Country	Medals	Country	Medals
Germany	29	Finland	9
United States	25	Czech Republic	4
Canada	24	Estonia	3
Austria	23	Croatia	3
Russia	22	Australia	2
Norway	19	Poland	2
Sweden	14	Ukraine	2
Switzerland	14	Japan	1
South Korea	11	Belarus	1
Italy	11	Bulgaria	1
China	11	Great Britain	1
France	9	Slovakia	1
Netherlands	9	Latvia	1

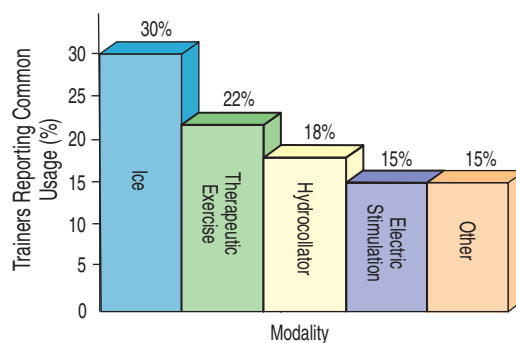
- Try to make a display of these data. What problems do you encounter?
- Can you find a way to organize the data so that the graph is more successful?

15. **Global Warming.** The Pew Research Center for the People and the Press (<http://people-press.org>) has asked a representative sample of U.S. adults about global warming, repeating the question over time. In January 2007, the responses reflected an increased belief that global warming is real and due to human activity. Here's a display of the percentages of respondents choosing each of the major alternatives offered:



List the errors in this display.

16. **Modalities.** A survey of athletic trainers (Scott F. Nadler, Michael Prybicien, Gerard A. Malanga, and Dan Sicher. "Complications from Therapeutic Modalities: Results of a National Survey of Athletic Trainers." *Archives of Physical Medical Rehabilitation* 84 [June 2003]) asked what modalities (treatment methods such as ice, whirlpool, ultrasound, or exercise) they commonly use to treat injuries. Respondents were each asked to list three modalities. The article included the following figure reporting the modalities used:



- What problems do you see with the graph?
- Consider the percentages for the named modalities. Do you see anything odd about them?

17. **Teen smokers.** The organization Monitoring the Future ([www.monitoringthefuture.org](http://www.monitoringthefuture.org)) asked 2048 eighth graders who said they smoked cigarettes what brands they preferred. The table below shows brand preferences for two regions of the country. Write a few sentences describing the similarities and differences in brand preferences among eighth graders in the two regions listed.

Brand preference	South	West
Marlboro	58.4%	58.0%
Newport	22.5%	10.1%
Camel	3.3%	9.5%
Other (over 20 brands)	9.1%	9.5%
No usual brand	6.7%	12.9%

18. **Handguns.** In an effort to reduce the number of gun-related homicides, some cities have run buyback programs in which the police offer cash (often \$50) to anyone who turns in an operating handgun. *Chance* magazine looked at results from a four-year period in Milwaukee. The table on the next page shows what types of guns were turned in and what types were used in homicides during a four-year period. Write a few sentences comparing the two distributions.

Caliber of gun	Buyback	Homicide
Small (.22, .25, .32)	76.4%	20.3%
Medium (.357, .38, 9 mm)	19.3%	54.7%
Large (.40, .44, .45)	2.1%	10.8%
Other	2.2%	14.2%

**T 19. Movies by Genre and Rating.** Here's a table that classifies movies released in 2005 by genre and MPAA rating:

	G	PG	PG-13	R	Total
Action/Adventure	66.7	25	30.4	23.7	29.2
Comedy	33.3	60.0	35.7	10.5	31.7
Drama	0	15.0	14.3	44.7	23.3
Thriller/Horror	0	0	19.6	21.1	15.8
Total	100%	100%	100%	100%	100%

- The table gives column percents. How could you tell that from the table itself?
- What percentage of these movies were comedies?
- What percentage of the PG-rated movies were comedies?
- Which of the following can you learn from this table? Give the answer if you can find it from the table.
  - The percentage of PG-13 movies that were comedies
  - The percentage of dramas that were R-rated
  - The percentage of dramas that were G-rated
  - The percentage of 2005 movies that were PG-rated comedies

**T 20. The Last Picture Show.** Here's another table showing information about 120 movies released in 2005. This table gives percentages of the table total:

	G	PG	PG-13	R	Total
Action/Adventure	3.33%	4.17	14.2	7.50	29.2
Comedy	1.67	10	16.7	3.33	31.7
Drama	0	2.50	6.67	14.2	23.3
Thriller/Horror	0	0	9.17	6.67	15.8
Total	5	16.7	46.7	31.7	100%

- How can you tell that this table holds table percentages (rather than row or column percentages)?
  - What was the most common genre/rating combination in 2005 movies?
  - How many of these movies were PG-rated comedies?
  - How many were G-rated?
  - An editorial about the movies noted, "More than three-quarters of the movies made today can be seen only by patrons 13 years old or older." Does this table support that assertion? Explain.
- 21. Seniors.** Prior to graduation, a high school class was surveyed about its plans. The following table displays the results for white and minority students (the "Minority"

group included African-American, Asian, Hispanic, and Native American students):

Seniors		
	White	Minority
4-year college	198	44
2-year college	36	6
Military	4	1
Employment	14	3
Other	16	3

- What percent of the seniors are white?
  - What percent of the seniors are planning to attend a 2-year college?
  - What percent of the seniors are white and planning to attend a 2-year college?
  - What percent of the white seniors are planning to attend a 2-year college?
  - What percent of the seniors planning to attend a 2-year college are white?
- 22. Politics.** Students in an Intro Stats course were asked to describe their politics as "Liberal," "Moderate," or "Conservative." Here are the results:

Politics					
		L	M	C	Total
Sex	Female	35	36	6	77
	Male	50	44	21	115
	Total	85	80	27	192

- What percent of the class is male?
  - What percent of the class considers themselves to be "Conservative"?
  - What percent of the males in the class consider themselves to be "Conservative"?
  - What percent of all students in the class are males who consider themselves to be "Conservative"?
- 23. More about seniors.** Look again at the table of post-graduation plans for the senior class in Exercise 21.
- Find the conditional distributions (percentages) of plans for the white students.
  - Find the conditional distributions (percentages) of plans for the minority students.
  - Create a graph comparing the plans of white and minority students.
  - Do you see any important differences in the post-graduation plans of white and minority students? Write a brief summary of what these data show, including comparisons of conditional distributions.
- 24. Politics revisited.** Look again at the table of political views for the Intro Stats students in Exercise 22.
- Find the conditional distributions (percentages) of political views for the females.
  - Find the conditional distributions (percentages) of political views for the males.
  - Make a graphical display that compares the two distributions.
  - Do the variables *Politics* and *Sex* appear to be independent? Explain.

25. **Magnet schools revisited.** The *Chance* magazine article described in Exercise 9 further examined the impact of an applicant's ethnicity on the likelihood of admission to the Houston Independent School District's magnet schools programs. Those data are summarized in the table below:

		Admission Decision			Total
		Accepted	Wait-listed	Turned away	
Ethnicity	Black/Hispanic	485	0	32	517
	Asian	110	49	133	292
	White	336	251	359	946
	Total	931	300	524	1755

- a) What percent of all applicants were Asian?  
 b) What percent of the students accepted were Asian?  
 c) What percent of Asians were accepted?  
 d) What percent of all students were accepted?
26. **More politics.** Look once more at the table summarizing the political views of Intro Stats students in Exercise 22.
- a) Produce a graphical display comparing the conditional distributions of males and females among the three categories of politics.  
 b) Comment briefly on what you see from the display in a.
27. **Back to school.** Examine the table about ethnicity and acceptance for the Houston Independent School District's magnet schools program, shown in Exercise 25. Does it appear that the admissions decisions are made independent of the applicant's ethnicity? Explain.
28. **Cars.** A survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table.

		Driver	
		Student	Staff
Origin	American	107	105
	European	33	12
	Asian	55	47

- a) What percent of all the cars surveyed were foreign?  
 b) What percent of the American cars were owned by students?  
 c) What percent of the students owned American cars?  
 d) What is the marginal distribution of origin?  
 e) What are the conditional distributions of origin by driver classification?  
 f) Do you think that the origin of the car is independent of the type of driver? Explain.
29. **Weather forecasts.** Just how accurate are the weather forecasts we hear every day? The following table compares the daily forecast with a city's actual weather for a year:

		Actual Weather	
		Rain	No rain
Forecast	Rain	27	63
	No rain	7	268

- a) On what percent of days did it actually rain?  
 b) On what percent of days was rain predicted?  
 c) What percent of the time was the forecast correct?  
 d) Do you see evidence of an association between the type of weather and the ability of forecasters to make an accurate prediction? Write a brief explanation, including an appropriate graph.
30. **Twins.** In 2000, the *Journal of the American Medical Association (JAMA)* published a study that examined pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without procedures, or term/post-term. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]:335–341)

		TWIN BIRTHS 1995–1997 (IN THOUSANDS)			Total
		Preterm (induced or cesarean)	Preterm (without procedures)	Term or post-term	
Level of Prenatal Care	Intensive	18	15	28	61
	Adequate	46	43	65	154
	Inadequate	12	13	38	63
	Total	76	71	131	278

- a) What percent of these mothers received inadequate medical care during their pregnancies?  
 b) What percent of all twin births were preterm?  
 c) Among the mothers who received inadequate medical care, what percent of the twin births were preterm?  
 d) Create an appropriate graph comparing the outcomes of these pregnancies by the level of medical care the mother received.  
 e) Write a few sentences describing the association between these two variables.
31. **Blood pressure.** A company held a blood pressure screening clinic for its employees. The results are summarized in the table below by age group and blood pressure level:

		Age		
		Under 30	30–49	Over 50
Blood Pressure	Low	27	37	31
	Normal	48	91	93
	High	23	51	73

- a) Find the marginal distribution of blood pressure level.
- b) Find the conditional distribution of blood pressure level within each age group.
- c) Compare these distributions with a segmented bar graph.
- d) Write a brief description of the association between age and blood pressure among these employees.
- e) Does this prove that people’s blood pressure increases as they age? Explain.

32. **Obesity and exercise.** The Centers for Disease Control and Prevention (CDC) has estimated that 19.8% of Americans over 15 years old are obese. The CDC conducts a survey on obesity and various behaviors. Here is a table on self-reported exercise classified by body mass index (BMI):

		Body Mass Index		
		Normal (%)	Overweight (%)	Obese (%)
Physical Activity	Inactive	23.8	26.0	35.6
	Irregularly active	27.8	28.7	28.1
	Regular, not intense	31.6	31.1	27.2
	Regular, intense	16.8	14.2	9.1

- a) Are these percentages column percentages, row percentages, or table percentages?
- b) Use graphical displays to show different percentages of physical activities for the three BMI groups.
- c) Do these data prove that lack of exercise causes obesity? Explain.

33. **Anorexia.** Hearing anecdotal reports that some patients undergoing treatment for the eating disorder anorexia seemed to be responding positively to the antidepressant Prozac, medical researchers conducted an experiment to investigate. They found 93 women being treated for anorexia who volunteered to participate. For one year, 49 randomly selected patients were treated with Prozac and the other 44 were given an inert substance called a placebo. At the end of the year, patients were diagnosed as healthy or relapsed, as summarized in the table:

	Prozac	Placebo	Total
Healthy	35	32	67
Relapse	14	12	26
Total	49	44	93

Do these results provide evidence that Prozac might be helpful in treating anorexia? Explain.

34. **Antidepressants and bone fractures.** For a period of five years, physicians at McGill University Health Center followed more than 5000 adults over the age of 50. The

researchers were investigating whether people taking a certain class of antidepressants (SSRIs) might be at greater risk of bone fractures. Their observations are summarized in the table:

	Taking SSRI	No SSRI	Total
Experienced fractures	14	244	258
No fractures	123	4627	4750
Total	137	4871	5008

Do these results suggest there’s an association between taking SSRI antidepressants and experiencing bone fractures? Explain.

35. **Drivers’ licenses 2005.** The following table shows the number of licensed U.S. drivers by age and by sex ([www.dot.gov](http://www.dot.gov)):

Age	Male Drivers (number)	Female Drivers (number)	Total
19 and under	4,777,694	4,553,946	9,331,640
20–24	8,611,161	8,398,879	17,010,040
25–29	8,879,476	8,666,701	17,546,177
30–34	9,262,713	8,997,662	18,260,375
35–39	9,848,050	9,576,301	19,424,351
40–44	10,617,456	10,484,149	21,101,605
45–49	10,492,876	10,482,479	20,975,355
50–54	9,420,619	9,475,882	18,896,501
55–59	8,218,264	8,265,775	16,484,039
60–64	6,103,732	6,147,569	12,251,361
65–69	4,571,157	4,643,913	9,215,070
70–74	3,617,908	3,761,039	7,378,947
75–79	2,890,155	3,192,408	6,082,563
80–84	1,907,743	2,222,412	4,130,155
85 and over	1,170,817	1,406,271	2,577,088
Total	100,389,881	100,275,386	200,665,267

- a) What percent of total drivers are under 20?
- b) What percent of total drivers are male?
- c) Write a few sentences comparing the number of male and female licensed drivers in each age group.
- d) Do a driver’s age and sex appear to be independent? Explain?

36. **Tattoos.** A study by the University of Texas Southwestern Medical Center examined 626 people to see if an increased risk of contracting hepatitis C was associated with having a tattoo. If the subject had a tattoo, researchers asked whether it had been done in a commercial tattoo parlor or elsewhere. Write a brief description of the association between tattooing and hepatitis C, including an appropriate graphical display.

	Tattoo done in commercial parlor	Tattoo done elsewhere	No tattoo
Has hepatitis C	17	8	18
No hepatitis C	35	53	495

37. **Hospitals.** Most patients who undergo surgery make routine recoveries and are discharged as planned. Others suffer excessive bleeding, infection, or other postsurgical complications and have their discharges from the hospital delayed. Suppose your city has a large hospital and a small hospital, each performing major and minor surgeries. You collect data to see how many surgical patients have their discharges delayed by postsurgical complications, and you find the results shown in the following table.

	Discharge Delayed	
	Large hospital	Small hospital
Major surgery	120 of 800	10 of 50
Minor surgery	10 of 200	20 of 250

- Overall, for what percent of patients was discharge delayed?
  - Were the percentages different for major and minor surgery?
  - Overall, what were the discharge delay rates at each hospital?
  - What were the delay rates at each hospital for each kind of surgery?
  - The small hospital advertises that it has a lower rate of postsurgical complications. Do you agree?
  - Explain, in your own words, why this confusion occurs.
38. **Delivery service.** A company must decide which of two delivery services it will contract with. During a recent trial period, the company shipped numerous packages with each service and kept track of how often deliveries did not arrive on time. Here are the data:

Delivery Service	Type of Service	Number of Deliveries	Number of Late Packages
Pack Rats	Regular	400	12
	Overnight	100	16
Boxes R Us	Regular	100	2
	Overnight	400	28

- Compare the two services' overall percentage of late deliveries.
- On the basis of the results in part a, the company has decided to hire Pack Rats. Do you agree that Pack Rats delivers on time more often? Explain.
- The results here are an instance of what phenomenon?

39. **Graduate admissions.** A 1975 article in the magazine *Science* examined the graduate admissions process at Berkeley for evidence of sex discrimination. The table below shows the number of applicants accepted to each of four graduate programs:

		Males accepted (of applicants)	Females accepted (of applicants)
Program	1	511 of 825	89 of 108
	2	352 of 560	17 of 25
	3	137 of 407	132 of 375
	4	22 of 373	24 of 341
	Total	1022 of 2165	262 of 849

- What percent of total applicants were admitted?
  - Overall, was a higher percentage of males or females admitted?
  - Compare the percentage of males and females admitted in each program.
  - Which of the comparisons you made do you consider to be the most valid? Why?
40. **Be a Simpson!** Can you design a Simpson's paradox? Two companies are vying for a city's "Best Local Employer" award, to be given to the company most committed to hiring local residents. Although both employers hired 300 new people in the past year, Company A brags that it deserves the award because 70% of its new jobs went to local residents, compared to only 60% for Company B. Company B concedes that those percentages are correct, but points out that most of its new jobs were full-time, while most of Company A's were part-time. Not only that, says Company B, but a higher percentage of its full-time jobs went to local residents than did Company A's, and the same was true for part-time jobs. Thus, Company B argues, it's a better local employer than Company A.
- Show how it's possible for Company B to fill a higher percentage of both full-time and part-time jobs with local residents, even though Company A hired more local residents overall.



### JUST CHECKING Answers

- 50.0%
- 44.4%
- 25.0%
- 15.6% Blue, 56.3% Brown, 28.1% Green/Hazel/Other
- 18.8% Blue, 62.5% Brown, 18.8% Green/Hazel/Other
- 40% of the blue-eyed students are female, while 50% of all students are female.
- Since blue-eyed students appear less likely to be female, it seems that *Sex* and *Eye Color* may not be independent. (But the numbers are small.)

# Displaying and Summarizing Quantitative Data



**T**sunamis are potentially destructive waves that can occur when the sea floor is suddenly and abruptly deformed. They are most often caused by earthquakes beneath the sea that shift the earth's crust, displacing a large mass of water.

The tsunami of December 26, 2004, with epicenter off the west coast of Sumatra, was caused by an earthquake of magnitude 9.0 on the Richter scale. It killed an estimated 297,248 people, making it the most disastrous tsunami on record. But was the earthquake that caused it truly extraordinary, or did it just happen at an unlucky place and time? The U.S. National Geophysical Data Center<sup>1</sup> has information on more than 2400 tsunamis dating back to 2000 B.C.E., and we have estimates of the magnitude of the underlying earthquake for 1240 of them. What can we learn from these data?

## Histograms

**WHO** 1240 earthquakes known to have caused tsunamis for which we have data or good estimates

**WHAT** Magnitude (Richter scale <sup>2</sup>), depth (m), date, location, and other variables

**WHEN** From 2000 B.C.E. to the present

**WHERE** All over the earth

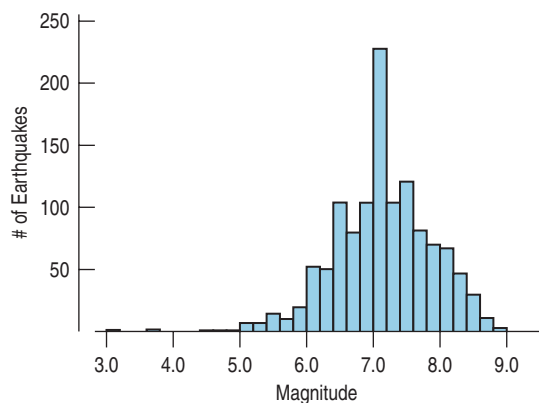
Let's start with a picture. For categorical variables, it is easy to draw the distribution because each category is a natural "pile." But for quantitative variables, there's no obvious way to choose piles. So, usually, we slice up all the possible values into equal-width bins. We then count the number of cases that fall into each bin. The bins, together with these counts, give the **distribution** of the quantitative variable and provide the building blocks for the histogram. By representing the counts as bars and plotting them against the bin values, the **histogram** displays the distribution at a glance.

<sup>1</sup> [www.ngdc.noaa.gov](http://www.ngdc.noaa.gov)

<sup>2</sup> Technically, Richter scale values are in units of log dyne-cm. But the Richter scale is so common now that usually the units are assumed. The U.S. Geological Survey gives the background details of Richter scale measurements on its Web site [www.usgs.gov/](http://www.usgs.gov/).



For example, here are the *Magnitudes* (on the Richter scale) of the 1240 earthquakes in the NGDC data:



**FIGURE 4.1**

A histogram of earthquake magnitudes shows the number of earthquakes with magnitudes (in Richter scale units) in each bin.

One surprising feature of the earthquake magnitudes is the spike around magnitude 7.0. Only one other bin holds even half that many earthquakes. These values include historical data for which the magnitudes were estimated by experts and not measured by modern seismographs. Perhaps the experts thought 7 was a typical and reasonable value for a tsunami-causing earthquake when they lacked detailed information. That would explain the overabundance of magnitudes right at 7.0 rather than spread out near that value.

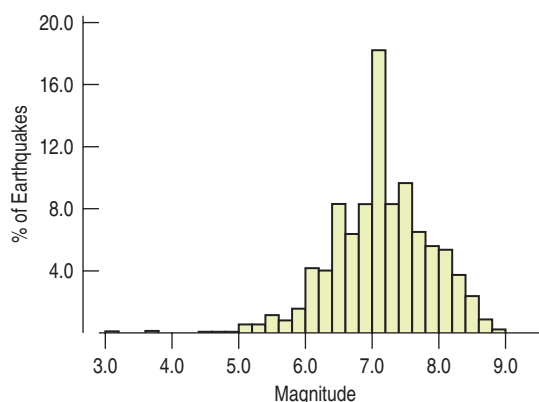
Like a bar chart, a histogram plots the bin counts as the heights of bars. In this histogram of earthquake magnitudes, each bin has a width of 0.2, so, for example, the height of the tallest bar says that there were about 230 earthquakes with magnitudes between 7.0 and 7.2. In this way, the histogram displays the entire distribution of earthquake magnitudes.

Does the distribution look as you expected? It is often a good idea to *imagine* what the distribution might look like before you make the display. That way you'll be less likely to be fooled by errors in the data or when you accidentally graph the wrong variable.

From the histogram, we can see that these earthquakes typically have magnitudes around 7. Most are between 5.5 and 8.5, and some are as small as 3 and as big as 9. Now we can answer the question about the Sumatra tsunami. With a value of 9.0 it's clear that the earthquake that caused it was an extraordinarily powerful earthquake—one of the largest on record.<sup>3</sup>

The bar charts of categorical variables we saw in Chapter 3 had spaces between the bars to separate the counts of different categories. But in a histogram, the bins slice up *all the values* of the quantitative variable, so any spaces in a histogram are actual **gaps** in the data, indicating a region where there are no values.

Sometimes it is useful to make a **relative frequency histogram**, replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the vertical scale is different.



**FIGURE 4.2**

A relative frequency histogram looks just like a frequency histogram except for the labels on the y-axis, which now show the percentage of earthquakes in each bin.

<sup>3</sup> Some experts now estimate the magnitude at between 9.1 and 9.3.

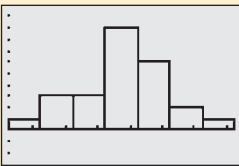
## T1 Tips

## Making a histogram

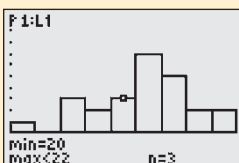
L1	L2	L3	1
22			
17			
18			
29			
22			
22			
23			
24			
23			
17			
21			
25			
20			
L1 = {22, 17, 18, 29...			

STAT	PLOTS
1: Plot1	On
2: Plot2	Off
3: Plot3	Off
4: Plots	Off

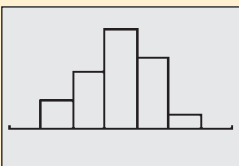
2nd	STAT	PLLOT	1
On	Off	Off	
Type:	Bar	Line	
Xlist:	L1		
Freq:	1		



WINDOW
Xmin=12
Xmax=30
Xscl=2
Ymin=-2.70621
Ymax=10.53
Yscl=1
Xres=3



L1	L2	L3	3
22	60		
17	70		
18	80		
29	90		
22	100		
22			
23			
L3{6} =			



Your calculator can create histograms. First you need some data. For an agility test, fourth-grade children jump from side to side across a set of parallel lines, counting the number of lines they clear in 30 seconds. Here are their scores:

22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21, 25, 20  
12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data into **L1**.

Now set up the calculator's plot:

- Go to **2nd STATPLOT**, choose **Plot1**, then **ENTER**.
- In the **Plot1** screen choose **On**, select the little histogram icon, then specify **Xlist:L1** and **Freq:1**.
- Be sure to turn off any other graphs the calculator may be set up for. Just hit the **Y=** button, and deactivate any functions seen there.

All set? To create your preliminary plot go to **ZOOM**, select **9:ZoomStat**, and then **ENTER**.

You now see the calculator's initial attempt to create a histogram of these data. Not bad. We can see that the distribution is roughly symmetric. But it's hard to tell exactly what this histogram shows, right? Let's fix it up a bit.

- Under **WINDOW**, let's reset the bins to convenient, sensible values. Try **Xmin=12**, **Xmax=30** and **Xscl=2**. That specifies the range of values along the *x*-axis and makes each bar span two lines.
- Hit **GRAPH** (not **ZoomStat**—this time we want control of the scale!).

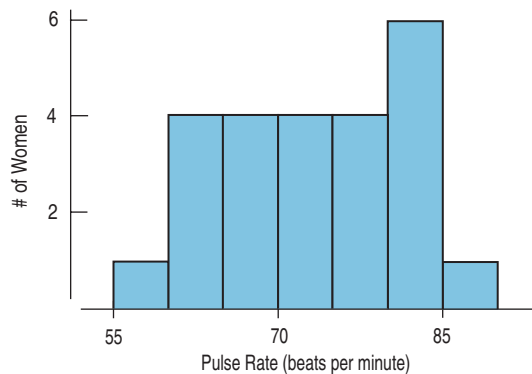
There. We still see rough symmetry, but also see that one of the scores was much lower than the others. Note that you can now find out exactly what the bars indicate by activating **TRACE** and then moving across the histogram using the arrow keys. For each bar the calculator will indicate the interval of values and the number of data values in that bin. We see that 3 kids had agility scores of 20 or 21.

Play around with the **WINDOW** settings. A different **Ymax** will make the bars appear shorter or taller. What happens if you set the bar width (**Xscl**) smaller? Or larger? You don't want to lump lots of values into just a few bins or make so many bins that the overall shape of the histogram is not clear. Choosing the best bar width takes practice.

Finally, suppose the data are given as a frequency table. Consider a set of test scores, with two grades in the 60s, four in the 70s, seven in the 80s, five in the 90s, and one 100. Enter the group cutoffs 60, 70, 80, 90, 100 in **L2** and the corresponding frequencies 2, 4, 7, 5, 1 in **L3**. When you set up the histogram **STATPLOT**, specify **Xlist:L2** and **Freq:L3**. Can you specify the **WINDOW** settings to make this histogram look the way you want it? (By the way, if you get a **DIM MISMATCH** error, it means you can't count. Look at **L2** and **L3**; you'll see the two lists don't have the same number of entries. Fix the problem by correcting the data you entered.)

## Stem-and-Leaf Displays

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. Here's a histogram of the pulse rates of 24 women, taken by a researcher at a health clinic:



**FIGURE 4.3**  
The pulse rates of 24 women at a health clinic

The Stem-and-Leaf display was devised by John W. Tukey, one of the greatest statisticians of the 20th century. It is called a "Stemplot" in some texts and computer programs, but we prefer Tukey's original name for it.

The story seems pretty clear. We can see the entire span of the data and can easily see what a typical pulse rate might be. But is that all there is to these data?

A **stem-and-leaf display** is like a histogram, but it shows the individual values. It's also easier to make by hand. Here's a stem-and-leaf display of the same data:

```

8 | 8
8 | 000044
7 | 6666
7 | 2222
6 | 8888
6 | 0444
5 | 6
Pulse Rate
(8|8 means 88 beats/min)

```

**AS** **Activity: Stem-and-Leaf Displays.** As you might expect of something called "stem-and-leaf," these displays grow as you consider each data value.

Turn the stem-and-leaf on its side (or turn your head to the right) and squint at it. It should look roughly like the histogram of the same data. Does it? Well, it's backwards because now the higher values are on the left, but other than that, it has the same shape.<sup>4</sup>

What does the line at the top of the display that says 8 | 8 mean? It stands for a pulse of 88 beats per minute (bpm). We've taken the tens place of the number and made that the "stem." Then we sliced off the ones place and made it a "leaf." The next line down is 8 | 000044. That shows that there were four pulse rates of 80 and two of 84 bpm.

Stem-and-leaf displays are especially useful when you make them by hand for batches of fewer than a few hundred data values. They are a quick way to display—and even to record—numbers. Because the leaves show the individual values, we can sometimes see even more in the data than the distribution's shape. Take another look at all the leaves of the pulse data. See anything

<sup>4</sup> You could make the stem-and-leaf with the higher values on the bottom. Usually, though, higher on the top makes sense.

unusual? At a glance you can see that they are all even. With a bit more thought you can see that they are all multiples of 4—something you couldn't possibly see from a histogram. How do you think the nurse took these pulses? Counting beats for a full minute or counting for only 15 seconds and multiplying by 4?

**How do stem-and-leaf displays work?** Stem-and-leaf displays work like histograms, but they show more information. They use part of the number itself (called the stem) to name the bins. To make the “bars,” they use the next digit of the number. For example, if we had a test score of 83, we could write it 8|3, where 8 serves as the stem and 3 as the leaf. Then, to display the scores 83, 76, and 88 together, we would write

$$\begin{array}{r|l} 8 & 38 \\ 7 & 6 \end{array}$$

For the pulse data, we have

$$\begin{array}{r|l} 8 & 0000448 \\ 7 & 22226666 \\ 6 & 04448888 \\ 5 & 6 \\ \text{Pulse Rate} & \\ (5|6 \text{ means } 56 \text{ beats/min}) & \end{array}$$

This display is OK, but a little crowded. A histogram might split each line into two bars. With a stem-and-leaf, we can do the same by putting the leaves 0–4 on one line and 5–9 on another, as we saw above:

$$\begin{array}{r|l} 8 & 8 \\ 8 & 000044 \\ 7 & 6666 \\ 7 & 2222 \\ 6 & 8888 \\ 6 & 0444 \\ 5 & 6 \\ \text{Pulse Rate} & \\ (8|8 \text{ means } 88 \text{ beats/min}) & \end{array}$$

For numbers with three or more digits, you'll often decide to truncate (or round) the number to two places, using the first digit as the stem and the second as the leaf. So, if you had 432, 540, 571, and 638, you might display them as shown below with an indication that 6|3 means 630–639.

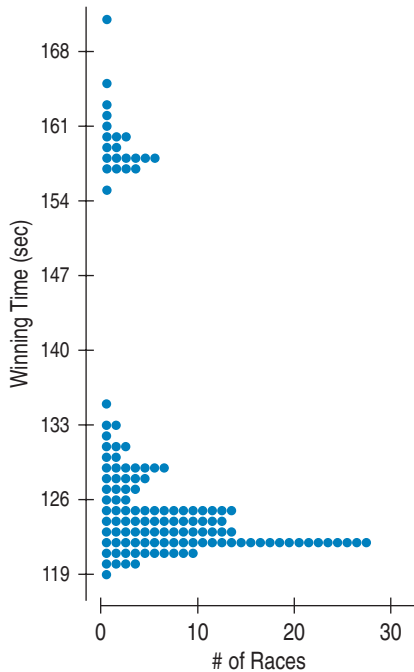
$$\begin{array}{r|l} 6 & 3 \\ 5 & 47 \\ 4 & 3 \end{array}$$

When you make a stem-and-leaf by hand, make sure to give each digit the same width, in order to preserve the area principle. (That can lead to some fat 1's and thin 8's—but it makes the display honest.)

## Dotplots

**AS**

**Activity: Dotplots.** Click on points to see their values and even drag them around.



A **dotplot** is a simple display. It just places a dot along an axis for each case in the data. It's like a stem-and-leaf display, but with dots instead of digits for all the leaves. Dotplots are a great way to display a small data set (especially if you forget how to write the digits from 0 to 9). Here's a dotplot of the time (in seconds) that the winning horse took to win the Kentucky Derby in each race between the first Derby in 1875 and the 2008 Derby.

Dotplots show basic facts about the distribution. We can find the slowest and quickest races by finding times for the topmost and bottommost dots. It's also clear that there are two clusters of points, one just below 160 seconds and the other at about 122 seconds. Something strange happened to the Derby times. Once we know to look for it, we can find out that in 1896 the distance of the Derby race was changed from 1.5 miles to the current 1.25 miles. That explains the two clusters of winning times.

Some dotplots stretch out horizontally, with the counts on the vertical axis, like a histogram. Others, such as the one shown here, run vertically, like a stem-and-leaf display. Some dotplots place points next to each other when they would otherwise overlap. Others just place them on top of one another. Newspapers sometimes offer dotplots with the dots made up of little pictures.

**FIGURE 4.4**

A dotplot of Kentucky Derby winning times plots each race as its own dot, showing the bimodal distribution.

## Think Before You Draw, Again

Suddenly, we face a lot more options when it's time to invoke our first rule of data analysis and make a picture. You'll need to *Think* carefully to decide which type of graph to make. In the previous chapter you learned to check the Categorical Data Condition before making a pie chart or a bar chart. Now, before making a stem-and-leaf display, a histogram, or a dotplot, you need to check the

**Quantitative Data Condition:** The data are values of a quantitative variable whose units are known.

Although a bar chart and a histogram may look somewhat similar, they're not the same display. You can't display categorical data in a histogram or quantitative data in a bar chart. Always check the condition that confirms what type of data you have before proceeding with your display.

Step back from a histogram or stem-and-leaf display. What can you say about the distribution? When you describe a distribution, you should always tell about three things: its **shape, center, and spread.**

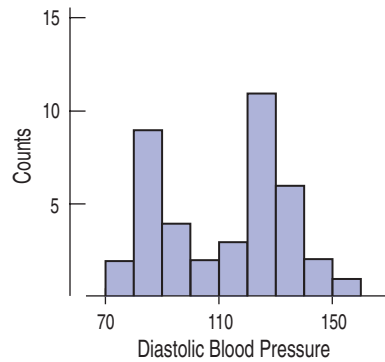
## The Shape of a Distribution

1. Does the histogram have a single, central hump or several separated humps? These humps are called **modes**.<sup>5</sup> The earthquake magnitudes have a single mode

<sup>5</sup> Well, technically, it's the value on the horizontal axis of the histogram that is the mode, but anyone asked to point to the mode would point to the hump.

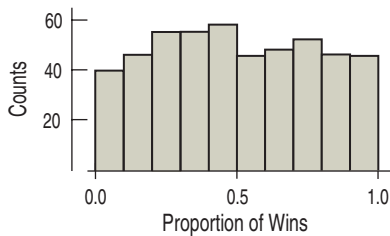
The **mode** is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. What is the mode of the Kentucky Derby times? Well, seven races were timed at 122.2 seconds—more than any other race time. Should that be the mode? Probably not. For quantitative data, it makes more sense to use the term “mode” in the more general sense of the peak of the histogram rather than as a single summary value. In this sense, the important feature of the Kentucky Derby races is that there are two distinct modes, representing the two different versions of the race and warning us to consider those two versions separately.

at just about 7. A histogram with one peak, such as the earthquake magnitudes, is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.<sup>6</sup> For example, here’s a bimodal histogram.



**FIGURE 4.5**  
A bimodal histogram has two apparent peaks.

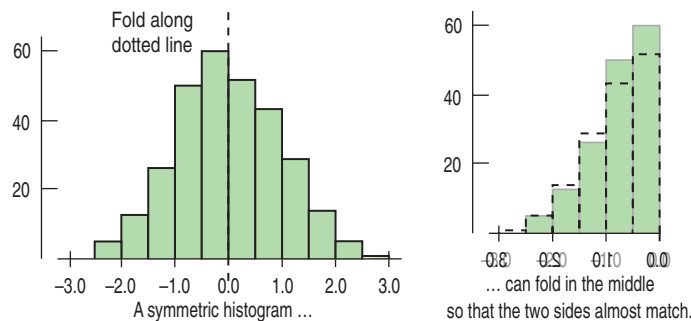
A histogram that doesn’t appear to have any mode and in which all the bars are approximately the same height is called **uniform**.



**FIGURE 4.6**  
In a uniform histogram, the bars are all about the same height. The histogram doesn’t appear to have a mode.

You’ve heard of pie à la mode. Is there a connection between pie and the mode of a distribution? Actually, there is! The mode of a distribution is a *popular* value near which a lot of the data values gather. And “à la mode” means “in style”—not “with ice cream.” That just happened to be a *popular* way to have pie in Paris around 1900.

2. *Is the histogram symmetric?* Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?

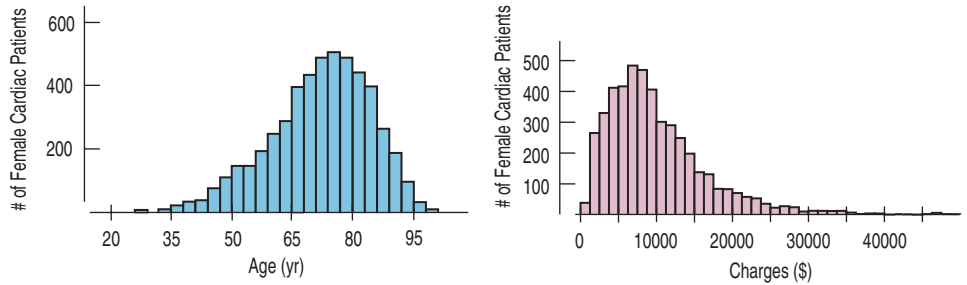


**FIGURE 4.7**

The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.

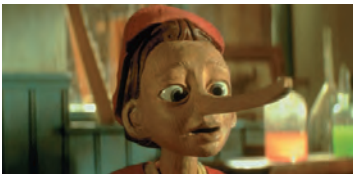
<sup>6</sup> Apparently, statisticians don’t like to count past two.

**AS** **Activity: Attributes of Distribution Shape.** This activity and the others on this page show off aspects of distribution shape through animation and example, then let you make and interpret histograms with your statistics package.



**FIGURE 4.8**

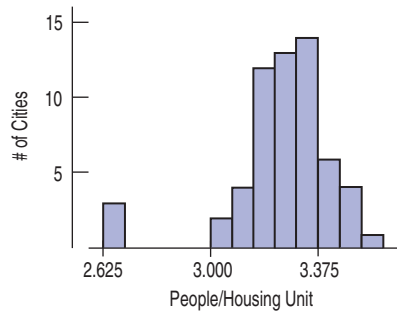
Two skewed histograms showing data on two variables for all female heart attack patients in New York state in one year. The blue one (age in years) is skewed to the left. The purple one (charges in \$) is skewed to the right.



3. *Do any unusual features stick out?* Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or outliers, that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it if you can. Be sure to look for outliers. Always.

In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.



**FIGURE 4.9**

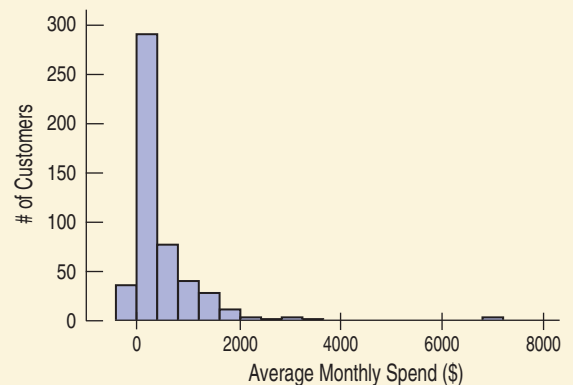
A histogram with outliers. There are three cities in the leftmost bar.

**FOR EXAMPLE** Describing histograms

A credit card company wants to see how much customers in a particular segment of their market use their credit card. They have provided you with data<sup>7</sup> on the amount spent by 500 selected customers during a 3-month period and have asked you to summarize the expenditures. Of course, you begin by making a histogram.

**Question:** Describe the shape of this distribution.

The distribution of expenditures is unimodal and skewed to the high end. There is an extraordinarily large value at about \$7000, and some of the expenditures are negative.



<sup>7</sup>These data are real, but cannot be further identified for obvious privacy reasons.

Are there any gaps in the distribution? The Kentucky Derby data that we saw in the dotplot on page 49 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes and encourage us to notice when the data may come from different sources or contain more than one group.



**Toto, I've a feeling we're not in math class anymore . . .** When Dorothy and her dog Toto land in Oz, everything is more vivid and colorful, but also more dangerous and exciting. Dorothy has new choices to make. She can't always rely on the old definitions, and the yellow brick road has many branches. You may be coming to a similar realization about Statistics.

When we summarize data, our goal is usually more than just developing a detailed knowledge of the data we have at hand. Scientists generally don't care about the particular guinea pigs they've treated, but rather about what their reactions say about how animals (and, perhaps, humans) would respond.

When you look at data, you want to know what the data say about the world, so you'd like to know whether the patterns you see in histograms and summary statistics generalize to other individuals and situations. You'll want to calculate summary statistics accurately, but then you'll also want to think about what they may say beyond just describing the data. And your knowledge about the world matters when you think about the overall meaning of your analysis.

It may surprise you that many of the most important concepts in Statistics are not defined as precisely as most concepts in mathematics. That's done on purpose, to leave room for judgment.

Because we want to see broader patterns rather than focus on the details of the data set we're looking at, we deliberately leave some statistical concepts a bit vague. Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the rest of the data to be considered an outlier—these are all somewhat vague concepts. And they all require judgment. You may be used to finding a single correct and precise answer, but in Statistics, there may be more than one interpretation. That may make you a little uncomfortable at first, but soon you'll see that this room for judgment brings you enormous power and responsibility. It means that using your own knowledge and judgment and supporting your findings with statistical evidence and justifications entitles you to your own opinions about what you see.



### JUST CHECKING

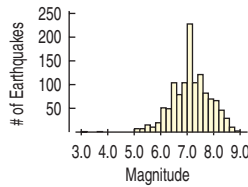
It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

## The Center of the Distribution: The Median

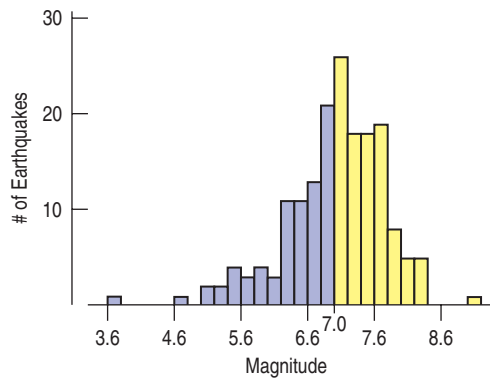
Let's return to the tsunami earthquakes. But this time, let's look at just 25 years of data: 176 earthquakes that occurred from 1981 through 2005. These should be more accurately measured than prehistoric quakes because seismographs were in wide use. Try to put your finger on the histogram at the value you think is





typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the **center** of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution such as these earthquake data, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed or possibly multimodal, it's not immediately clear what we even mean by the center.

One reasonable choice of typical value is the value that is literally in the middle, with half the values below it and half above it.



**FIGURE 4.10** *Tsunami-causing earthquakes (1981–2005)*

*The median splits the histogram into two halves of equal area.*

Histograms follow the area principle, and each half of the data has about 88 earthquakes, so each colored region has the same area in the display. The middle value that divides the histogram into two equal areas is called the **median**.

The median has the same units as the data. Be sure to include the units whenever you discuss the median.

For the recent tsunamis, there are 176 earthquakes, so the median is found at the  $(176 + 1)/2 = 88.5$ th place in the sorted data. That “.5” just says to average the two values on either side: the 88th and the 89th. The median earthquake magnitude is 7.0.

### NOTATION ALERT:

We always use  $n$  to indicate the number of values. Some people even say, “How big is the  $n$ ?” when they mean the number of data values.

**How do medians work?** Finding the median of a batch of  $n$  numbers is easy as long as you remember to order the values first. If  $n$  is odd, the median is the middle value. Counting in from the ends, we find this value in the  $\frac{n+1}{2}$  position.

When  $n$  is even, there are two middle values. So, in this case, the median is the average of the two values in positions  $\frac{n}{2}$  and  $\frac{n}{2} + 1$ .

Here are two examples:

Suppose the batch has these values: 14.1, 3.2, 25.3, 2.8,  $-17.5$ , 13.9, 45.8.

First we order the values:  $-17.5$ , 2.8, 3.2, 13.9, 14.1, 25.3, 45.8.

Since there are 7 values, the median is the  $(7 + 1)/2 = 4$ th value, counting from the top or bottom: 13.9. Notice that 3 values are lower, 3 higher.

Suppose we had the same batch with another value at 35.7. Then the ordered values are  $-17.5$ , 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, 45.8.

The median is the average of the  $8/2$  or 4th, and the  $(8/2) + 1$ , or 5th, values. So the median is  $(13.9 + 14.1)/2 = 14.0$ . Four data values are lower, and four higher.

The median is one way to find the center of the data. But there are many others. We'll look at an even more important measure later in this chapter.

Knowing the median, we could say that a typical tsunami-causing earthquake, worldwide, was about 7.0 on the Richter scale. How much does that really say? How well does the median describe the data? After all, not every earthquake has a Richter scale value of 7.0. Whenever we find the center of data, the next step is always to ask how well it actually summarizes the data.

## Spread: Home on the Range

Statistics pays close attention to what we *don't* know as well as what we do know. Understanding how spread out the data are is a first step in understanding what a summary *cannot* tell us about the data. It's the beginning of telling us what we don't know.

If every earthquake that caused a tsunami registered 7.0 on the Richter scale, then knowing the median would tell us everything about the distribution of earthquake magnitudes. The more the data vary, however, the less the median alone can tell us. So we need to measure how much the data values vary around the center. In other words, how spread out are they? **When we describe a distribution numerically, we always report a measure of its spread along with its center.**

How should we measure the spread? We could simply look at the extent of the data. How far apart are the two extremes? **The range of the data is defined as the difference between the maximum and minimum values:**

$$\text{Range} = \text{max} - \text{min}.$$

Notice that the range is a *single number*, not an interval of values, as you might think from its use in common speech. The maximum magnitude of these earthquakes is 9.0 and the minimum is 3.7, so the *range* is  $9.0 - 3.7 = 5.3$ .

The range has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall.

## Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. **One quarter of the data lies below the lower quartile, and one quarter of the data lies above the upper quartile, so half the data lies between them. The quartiles border the middle half of the data.**

**How do quartiles work?** A simple way to find the quartiles is to start by splitting the batch into two halves at the median. (When  $n$  is odd, some statisticians include the median in both halves; others omit it.) The lower quartile is the median of the lower half, and the upper quartile is the median of the upper half.

Here are our two examples again.

The ordered values of the first batch were  $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3,$  and  $45.8$ , with a median of  $13.9$ . Excluding the median, the two halves of the list are  $-17.5, 2.8, 3.2$  and  $14.1, 25.3, 45.8$ .

Each half has 3 values, so the median of each is the middle one. The lower quartile is  $2.8$ , and the upper quartile is  $25.3$ .

The second batch of data had the ordered values  $-17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7,$  and  $45.8$ .

Here  $n$  is even, so the two halves of 4 values are  $-17.5, 2.8, 3.2, 13.9$  and  $14.1, 25.3, 35.7, 45.8$ .

Now the lower quartile is  $(2.8 + 3.2)/2 = 3.0$ , and the upper quartile is  $(25.3 + 35.7)/2 = 30.5$ .

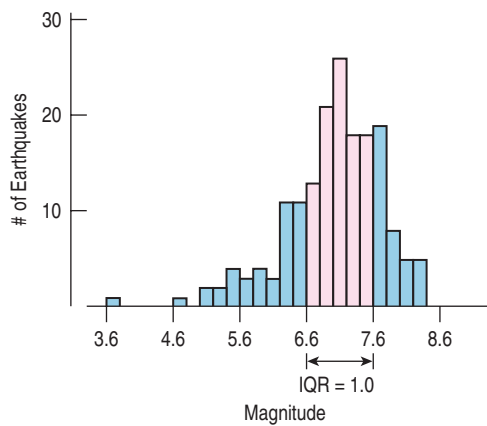
The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It's commonly abbreviated IQR (and pronounced "eye-cue-are," not "ikker"):

$$IQR = \text{upper quartile} - \text{lower quartile}.$$

For the earthquakes, there are 88 values below the median and 88 values above the median. The midpoint of the lower half is the average of the 44th and 45th values in the ordered data; that turns out to be 6.6. In the upper half we average the 132nd and 133rd values, finding a magnitude of 7.6 as the third quartile. The *difference* between the quartiles gives the IQR:

$$IQR = 7.6 - 6.6 = 1.0.$$

Now we know that the middle half of the earthquake magnitudes extends across a (interquartile) range of 1.0 Richter scale units. This seems like a reasonable summary of the spread of the distribution, as we can see from this histogram:



**FIGURE 4.11**

The quartiles bound the middle 50% of the values of the distribution. This gives a visual indication of the spread of the data. Here we see that the IQR is 1.0 Richter scale units.

The IQR is almost always a reasonable summary of the spread of a distribution. Even if the distribution itself is skewed or has some outliers, the IQR should provide useful information. The one exception is when the data are strongly bimodal. For example, remember the dotplot of winning times in the Kentucky Derby (page 49)? Because the race distance was changed, we really have data on two different races, and they shouldn't be summarized together.

**So, what is a quartile anyway?** Finding the quartiles sounds easy, but surprisingly, the quartiles are not well-defined. It's not always clear how to find a value such that exactly one quarter of the data lies above or below that value. We offered a simple rule for Finding Quartiles in the box on page 54: Find the median of each half of the data split by the median. When  $n$  is odd, we (and your TI calculator) omit the median from each of the halves. Some other texts include the median in both halves before finding the quartiles. Both methods are commonly used. If you are willing to do a bit more calculating, there are several other methods that locate a quartile somewhere between adjacent data values. We know of at least six different rules for finding quartiles. Remarkably, each one is in use in some software package or calculator.

So don't worry too much about getting the "exact" value for a quartile. All of the methods agree pretty closely when the data set is large. When the data set is small, different rules will disagree more, but in that case there's little need to summarize the data anyway.

Remember, Statistics is about understanding the world, not about calculating the right number. The "answer" to a statistical question is a sentence about the issue raised in the question.

The lower and upper quartiles are also known as the 25th and 75th percentiles of the data, respectively, since the lower quartile falls above 25% of the data and the upper quartile falls above 75% of the data. If we count this way, the median is the 50th percentile. We could, of course, define and calculate any percentile that we want. For example, the 10th percentile would be the number that falls above the lowest 10% of the data values.

## 5-Number Summary

### NOTATION ALERT:

We always use Q1 to label the lower (25%) quartile and Q3 to label the upper (75%) quartile. We skip the number 2 because the median would, by this system, naturally be labeled Q2—but we don't usually call it that.

The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). The 5-number summary for the recent tsunami earthquake *Magnitudes* looks like this:

Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

It's good idea to report the number of data values and the identity of the cases (the *Who*). Here there are 176 earthquakes.

The 5-number summary provides a good overview of the distribution of magnitudes of these tsunami-causing earthquakes. For a start, we can see that the median magnitude is 7.0. Because the IQR is only  $7.6 - 6.6 = 1$ , we see that many quakes are close to the median magnitude. Indeed, the quartiles show us that the middle half of these earthquakes had magnitudes between 6.6 and 7.6. One quarter of the earthquakes had magnitudes above 7.6, although one tsunami was caused by a quake measuring only 3.7 on the Richter scale.

### STEP-BY-STEP EXAMPLE

### Shape, Center, and Spread: Flight Cancellations



The U.S. Bureau of Transportation Statistics ([www.bts.gov](http://www.bts.gov)) reports data on airline flights. Let's look at data giving the percentage of flights cancelled each month between 1995 and 2005.

**Question:** How often are flights cancelled?

<b>WHO</b>	Months
<b>WHAT</b>	Percentage of flights cancelled at U.S. airports
<b>WHEN</b>	1995–2005
<b>WHERE</b>	United States



**Variable:** Identify the *variable*, and decide how you wish to display it.

To identify a variable, report the W's.

Select an appropriate display based on the nature of the data and what you want to know.

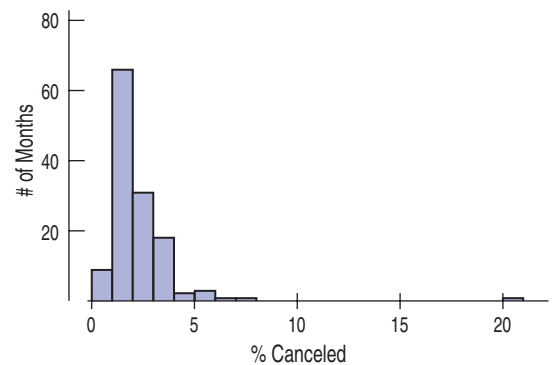
I want to learn about the monthly percentage of flight cancellations at U.S. airports.

I have data from the U.S. Bureau of Transportation Statistics giving the percentage of flights cancelled at U.S. airports each month between 1995 and 2005.

✓ **Quantitative Data Condition:** Percentages are quantitative. A histogram and numerical summaries would be appropriate.



**Mechanics:** We usually make histograms with a computer or graphing calculator.



The histogram shows a distribution skewed to the high end and one extreme outlier, a month in which more than 20% of flights were cancelled.

In most months, fewer than 5% of flights are cancelled and usually only about 2% or 3%. That seems reasonable.



It's always a good idea to think about what you expect to see so that you can check whether the histogram looks like what you expected.

With 132 cases, we probably have more data than you'd choose to work with by hand. The results given here are from technology.

Count	132
Max	20.240
Q3	2.615
Median	1.755
Q1	1.445
Min	0.770
IQR	1.170

TELL

**Interpretation:** Describe the shape, center, and spread of the distribution. Report on the symmetry, number of modes, and any gaps or outliers. You should also mention any concerns you may have about the data.

The distribution of cancellations is skewed to the right, and this makes sense: The values can't fall below 0%, but can increase almost arbitrarily due to bad weather or other events.

The median is 1.76% and the IQR is 1.17%. The low IQR indicates that in most months the cancellation rate is close to the median. In fact, it's between 1.4% and 2.6% in the middle 50% of all months, and in only 1/4 of the months were more than 2.6% of flights cancelled.

There is one extraordinary value: 20.2%. Looking it up, I find that the extraordinary month was September 2001. The attacks of September 11 shut down air travel for several days, accounting for this outlier.

## Summarizing Symmetric Distributions: The Mean

### NOTATION ALERT:

In Algebra you used letters to represent values in a problem, but it didn't matter what letter you picked. You could call the width of a rectangle  $X$  or you could call it  $w$  (or *Fred*, for that matter). But in Statistics, the notation is part of the vocabulary. For example, in Statistics  $n$  is always the number of data values. Always.

We have already begun to point out such special notation conventions:  $n$ ,  $Q1$ , and  $Q3$ . Think of them as part of the terminology you need to learn in this course.

Here's another one: Whenever we put a bar over a symbol, it means "find the mean."

Medians do a good job of summarizing the center of a distribution, even when the shape is skewed or when there is an outlier, as with the flight cancellations. But when we have symmetric data, there's another alternative. You probably already know how to average values. In fact, to find the median when  $n$  is even, we said you should average the two middle values, and you didn't even flinch.

The earthquake magnitudes are pretty close to symmetric, so we can also summarize their center with a mean. The mean tsunami earthquake magnitude is 6.96—about what we might expect from the histogram. You already know how to average values, but this is a good place to introduce notation that we'll use throughout the book. We use the Greek capital letter sigma,  $\Sigma$ , to mean "sum" (sigma is "S" in Greek), and we'll write:

$$\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}.$$

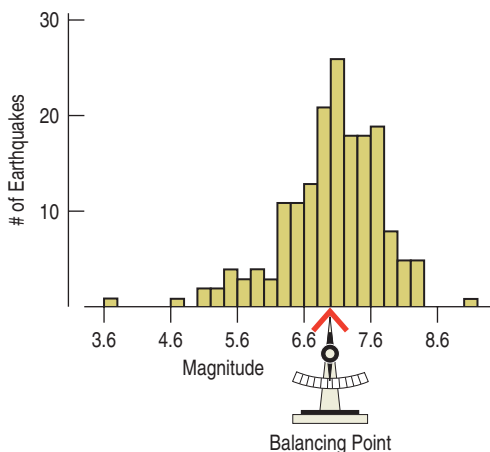
The formula says to add up all the values of the variable and divide that sum by the number of data values,  $n$ —just as you've always done.<sup>8</sup>

Once we've averaged the data, you'd expect the result to be called the *average*, but that would be too easy. Informally, we speak of the "average person" but we don't add up people and divide by the number of people. A median is also a kind of average. To make this distinction, the value we calculated is called the mean,  $\bar{y}$ , and pronounced "y-bar."

<sup>8</sup> You may also see the variable called  $x$  and the equation written  $\bar{x} = \frac{\text{Total}}{n} = \frac{\sum x}{n}$ . Don't let that throw you. You are free to name the variable anything you want, but we'll generally use  $y$  for variables like this that we want to summarize, model, or predict. (Later we'll talk about variables that are used to explain, model, or predict  $y$ . We'll call them  $x$ .)

The **mean** feels like the center because it is the point where the histogram balances:

In everyday language, sometimes “average” does mean what we want it to mean. We don’t talk about your grade point mean or a baseball player’s batting mean or the Dow Jones Industrial mean. So we’ll continue to say “average” when that seems most natural. When we do, though, you may assume that what we mean is the mean.

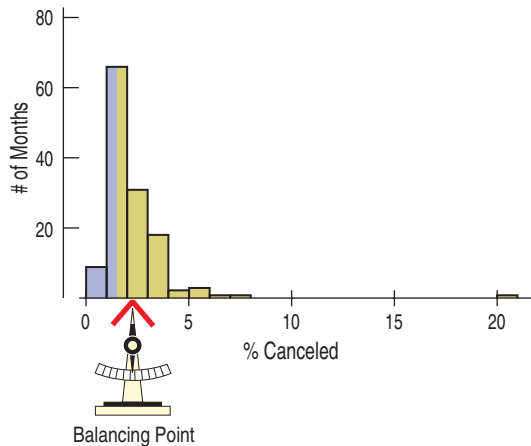


**FIGURE 4.12**  
The mean is located at the balancing point of the histogram.

## Mean or Median?

Using the center of balance makes sense when the data are symmetric. But data are not always this well behaved. If the distribution is skewed or has outliers, the center is not so well defined and the mean may not be what we want. For example, the mean of the flight cancellations doesn’t give a very good idea of the typical percentage of cancellations.

**TI-*nspire***  
**Mean, median, and outliers.**  
Drag data points around to explore how outliers affect the mean and median.



**FIGURE 4.13**  
The median splits the area of the histogram in half at 1.75%. Because the distribution is skewed to the right, the mean (2.28%) is higher than the median. The points at the right have pulled the mean toward them away from the median.

**A S** **Activity: The Center of a Distribution.** Compare measures of center by dragging points up and down and seeing the consequences. Another activity shows how to find summaries with your statistics package.

The mean is 2.28%, but nearly 70% of months had cancellation rates below that, so the mean doesn’t feel like a good overall summary. Why is the balancing point so high? The large outlying value pulls it to the right. For data like these, the median is a better summary of the center.

Because the median considers only the order of the values, it is **resistant** to values that are extraordinarily large or small; it simply notes that they are one of the “big ones” or the “small ones” and ignores their distance from the center.

For the tsunami earthquake magnitudes, it doesn’t seem to make much difference—the mean is 6.96; the median is 7.0. When the data are symmetric, the mean and median will be close, but when the data are skewed, the median is likely to be a better choice. So, why not just use the median? Well, for one, the median can go overboard. It’s not just resistant to occasional outliers, but can be unaffected by changes in up to half the data values. By contrast, the mean includes input from

each data value and gives each one equal weight. It's also easier to work with, so when the distribution is unimodal and symmetric, we'll use the mean.

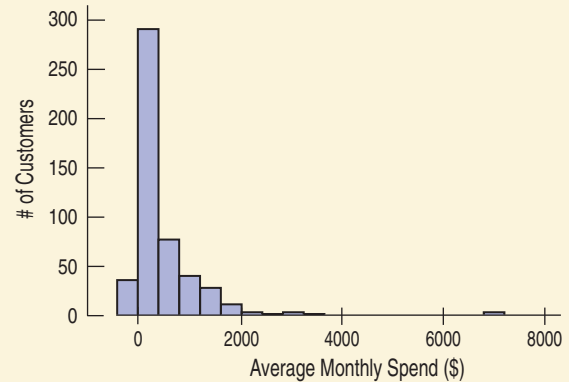
Of course, to choose between mean and median, we'll start by looking at the data. If the histogram is symmetric and there are no outliers, we'll prefer the mean. However, if the histogram is skewed or has outliers, we're usually better off with the median. If you're not sure, report both and discuss why they might differ.

## FOR EXAMPLE

### Describing center

**Recap:** You want to summarize the expenditures of 500 credit card company customers, and have looked at a histogram.

**Question:** You have found the mean expenditure to be \$478.19 and the median to be \$216.28. Which is the more appropriate measure of center, and why?



Because the distribution of expenditures is skewed, the median is the more appropriate measure of center. Unlike the mean, it's not affected by the large outlying value or by the skewness. Half of these credit card customers had average monthly expenditures less than \$216.28 and half more.

**When to expect skewness** Even without making a histogram, we can expect some variables to be skewed. When values of a quantitative variable are bounded on one side but not the other, the distribution may be skewed. For example, incomes and waiting times can't be less than zero, so they are often skewed to the right. Amounts of things (dollars, employees) are often skewed to the right for the same reason. If a test is too easy, the distribution will be skewed to the left because many scores will bump against 100%. And combinations of things are often skewed. In the case of the cancelled flights, flights are more likely to be cancelled in January (due to snowstorms) and in August (thunderstorms). Combining values across months leads to a skewed distribution.

## What About Spread? The Standard Deviation

**AS** **Activity: The Spread of a Distribution.** What happens to measures of spread when data values change may not be quite what you expect.

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation**, which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

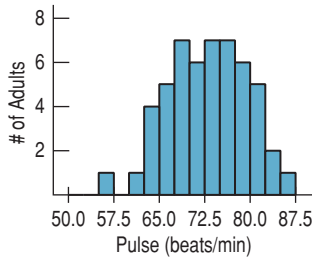
To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.



**NOTATION ALERT:**

$s^2$  always means the variance of a set of data, and  $s$  always denotes the standard deviation.

**WHO** 52 adults  
**WHAT** Resting heart rates  
**UNITS** Beats per minute



When we add up these squared deviations and find their average (almost), we call the result the **variance**:

$$s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Why almost? It *would* be a mean if we divided the sum by  $n$ . Instead, we divide by  $n - 1$ . Why? The simplest explanation is “to drive you crazy.” But there are good technical reasons, some of which we’ll see later.

The variance will play an important role later in this book, but it has a problem as a measure of spread. Whatever the units of the original data are, the variance is in *squared* units. We want measures of spread to have the same units as the data. And we probably don’t want to talk about squared dollars or *mpg*<sup>2</sup>. So, to get back to the original units, we take the square root of  $s^2$ . The result,  $s$ , is the **standard deviation**.

Putting it all together, the standard deviation of the data is found by the following formula:

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

You will almost always rely on a calculator or computer to do the calculating.

Understanding what the standard deviation really means will take some time, and we’ll revisit the concept in later chapters. For now, have a look at this histogram of resting pulse rates. The distribution is roughly symmetric, so it’s okay to choose the mean and standard deviation as our summaries of center and spread. The mean pulse rate is 72.7 beats per minute, and we can see that’s a typical heart rate. We also see that some heart rates are higher and some lower—but how much? Well, the standard deviation of 6.5 beats per minute indicates that, on average, we might expect people’s heart rates to differ from the mean rate by about 6.5 beats per minute. Looking at the histogram, we can see that 6.5 beats above or below the mean appears to be a typical deviation.

**How does standard deviation work?** To find the standard deviation, start with the mean,  $\bar{y}$ . Then find the *deviations* by taking  $\bar{y}$  from each value:  $(y - \bar{y})$ . Square each deviation:  $(y - \bar{y})^2$ .

Now you’re nearly home. Just add these up and divide by  $n - 1$ . That gives you the variance,  $s^2$ . To find the standard deviation,  $s$ , take the square root. Here we go:

Suppose the batch of values is 14, 13, 20, 22, 18, 19, and 13.

The mean is  $\bar{y} = 17$ . So the deviations are found by subtracting 17 from each value:

Original Values	Deviations	Squared Deviations
14	$14 - 17 = -3$	$(-3)^2 = 9$
13	$13 - 17 = -4$	$(-4)^2 = 16$
20	$20 - 17 = 3$	9
22	$22 - 17 = 5$	25
18	$18 - 17 = 1$	1
19	$19 - 17 = 2$	4
13	$13 - 17 = -4$	16

Add up the squared deviations:  $9 + 16 + 9 + 25 + 1 + 4 + 16 = 80$ .

Now divide by  $n - 1$ :  $80/6 = 13.33$ .

Finally, take the square root:  $s = \sqrt{13.33} = 3.65$

## Thinking About Variation

**A S** **Activity: Displaying Spread.** What does the standard deviation look like on a histogram? How about the IQR?

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? The average waiting time is the same. But the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

Statistics is about variation, so spread is an important fundamental concept in Statistics. Measures of spread help us to be precise about what we *don't* know. If many data values are scattered far from the center, the IQR and the standard deviation will be large. If the data values are close to the center, then these measures of spread will be small. If all our data values were exactly the same, we'd have no question about summarizing the center, and all measures of spread would be zero—and we wouldn't need Statistics. You might think this would be a big plus, but it would make for a boring world. Fortunately (at least for Statistics), data do vary.

Measures of spread tell how well other summaries describe the data. That's why we always (always!) report a spread along with any summary of the center.



### JUST CHECKING

- The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?
- You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?
- A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

## What to Tell About a Quantitative Variable

**A S** **Activity: Playing with Summaries.** Here's a Statistics game about summaries that even some experienced statisticians find . . . well, challenging. Your intuition may be better. Give it a try!

**TI-*n*spire**  
Standard deviation, IQR, and outliers. Drag data points around to explore how outliers affect measures of spread.

What should you *Tell* about a quantitative variable?

- ▶ Start by making a histogram or stem-and-leaf display, and discuss the shape of the distribution.
- ▶ Next, discuss the center *and* spread.
  - ▶ We always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.
  - ▶ If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation as well, but you should point out why the mean and median differ.
  - ▶ If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For unimodal symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true of your data set, look again to make sure that the distribution isn't skewed and there are no outliers.

**How “Accurate” Should We Be?**

Don’t think you should report means and standard deviations to a zillion decimal places; such implied accuracy is really meaningless. Although there is no ironclad rule, statisticians commonly report summary statistics to one or two decimal places more than the original data have.

- ▶ Also, discuss any unusual features.
  - ▶ If there are multiple modes, try to understand why. If you can identify a reason for separate modes (for example, women and men typically have heart attacks at different ages), it may be a good idea to split the data into separate groups.
  - ▶ If there are any clear outliers, you should point them out. If you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. (Of course, the median and IQR won’t be affected very much by the outliers.)

**STEP-BY-STEP EXAMPLE**

**Summarizing a distribution**

One of the authors owned a 1989 Nissan Maxima for 8 years. Being a statistician, he recorded the car’s fuel efficiency (in mpg) each time he filled the tank. He wanted to know what fuel efficiency to expect as “ordinary” for his car. (Hey, he’s a statistician. What would you expect?<sup>9</sup>) Knowing this, he was able to predict when he’d need to fill the tank again and to notice if the fuel efficiency suddenly got worse, which could be a sign of trouble.

**Question:** How would you describe the distribution of *Fuel efficiency* for this car?



**Plan** State what you want to find out.

**Variable** Identify the variable and report the W’s.

Be sure to check the appropriate condition.

I want to summarize the distribution of Nissan Maxima fuel efficiency.

The data are the fuel efficiency values in miles per gallon for the first 100 fill-ups of a 1989 Nissan Maxima between 1989 and 1992.

✓ **Quantitative Data Condition:** The fuel efficiencies are quantitative with units of miles per gallon. Histograms and boxplots are appropriate displays for displaying the distribution. Numerical summaries are appropriate as well.

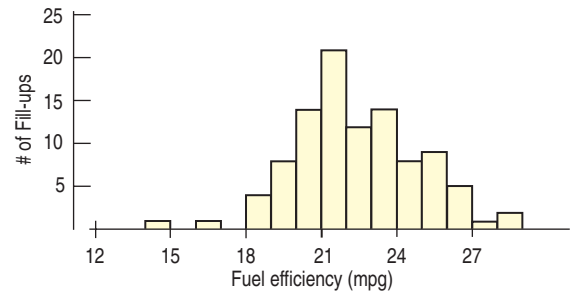
<sup>9</sup> He also recorded the time of day, temperature, price of gas, and phase of the moon. (OK, maybe not phase of the moon.) His data are on the DVD.

SHOW

**Mechanics** Make a histogram and boxplot. Based on the shape, choose appropriate numerical summaries.

REALITY CHECK

A value of 22 mpg seems reasonable for such a car. The spread is reasonable, although the range looks a bit large.



A histogram of the data shows a fairly symmetric distribution with a low outlier.

Count	100
Mean	22.4 mpg
StdDev	2.45
Q1	20.8
Median	22.0
Q3	24.0
IQR	3.2

The mean and median are close, so the outlier doesn't seem to be a problem. I can use the mean and standard deviation.

TELL

**Conclusion** Summarize and interpret your findings in context. Be sure to discuss the distribution's shape, center, spread, and unusual features (if any).

The distribution of mileage is unimodal and roughly symmetric with a mean of 22.4 mpg. There is a low outlier that should be investigated, but it does not influence the mean very much. The standard deviation suggests that from tankful to tankful, I can expect the car's fuel economy to differ from the mean by an average of about 2.45 mpg.

**Are my statistics "right"?** When you calculate a mean, the computation is clear: You sum all the values and divide by the sample size. You may round your answer less or more than someone else (we recommend one more decimal place than the data), but all books and technologies agree on how to find the mean. Some statistics, however, are more problematic. For example we've already pointed out that methods of finding quartiles differ.

Differences in numeric results can also arise from decisions in the middle of calculations. For example, if you round off your value for the mean before you calculate the sum of squared deviations, your standard deviation probably won't agree with a computer program that calculates using many decimal places. (We do recommend that you do calculations using as many digits as you can to minimize this effect.)

Don't be overly concerned with these discrepancies, especially if the differences are small. They don't mean that your answer is "wrong," and they usually won't change any conclusion you might draw about the data. Sometimes (in footnotes and in the answers in the back of the book) we'll note alternative results, but we could never list all the possible values, so we'll rely on your common sense to focus on the meaning rather than on the digits. Remember: Answers are sentences!

TI Tips

Calculating the statistics

```

EDIT [2nd] [MODE] TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
    
```

```

1-Var Stats L1
    
```

```

1-Var Stats
x=22
Σx=550
Σx²=12480
Sx=3.979112129
σx=3.898717738
n=25
    
```

```

1-Var Stats
n=25
minX=12
Q1=19.5
Med=22
Q3=25
maxX=29
    
```

Your calculator can easily find all the numerical summaries of data. To try it out, you simply need a set of values in one of your datalists. We'll illustrate using the boys' agility test results from this chapter's earlier TI Tips (still in L1), but you can use any data currently stored in your calculator.

- Under the **STAT** **CALC** menu, select **1-Var Stats** and hit **ENTER**.
- Specify the location of your data, creating a command like **1-Var Stats L1**.
- Hit **ENTER** again.

Voilà! Everything you wanted to know, and more. Among all of the information shown, you are primarily interested in these statistics:  $\bar{x}$  (the mean),  $S_x$  (the standard deviation),  $n$  (the count), and—scrolling down—**minX** (the smallest datum), **Q<sub>1</sub>** (the first quartile), **Med** (the median), **Q<sub>3</sub>** (the third quartile), and **maxX** (the largest datum).

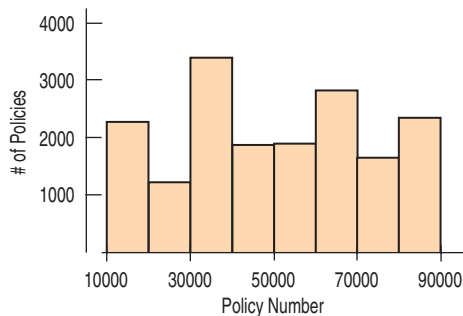
Sorry, but the TI doesn't explicitly tell you the range or the IQR. Just subtract:  $IQR = Q_3 - Q_1 = 25 - 19.5 = 5.5$ . What's the range?

By the way, if the data come as a frequency table with the values stored in, say, **L4** and the corresponding frequencies in **L5**, all you have to do is ask for **1-Var Stats L4,L5**.

**WHAT CAN GO WRONG?**

A data display should tell a story about the data. To do that, it must speak in a clear language, making plain what variable is displayed, what any axis shows, and what the values of the data are. And it must be consistent in those decisions.

A display of quantitative data can go wrong in many ways. The most common failures arise from only a few basic errors:



**FIGURE 4.14**  
It's not appropriate to display these data with a histogram.

▶ **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean that it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are just labels. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart would be more appropriate.

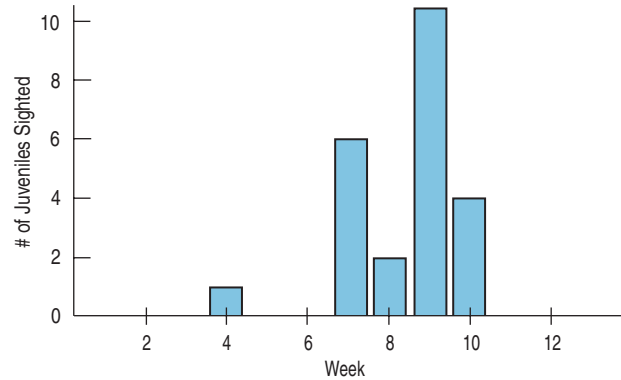
▶ **Don't look for shape, center, and spread of a bar chart.** A bar chart showing the sizes of the piles displays the distribution of a categorical variable, but the bars could be arranged in any order left to right. Concepts like symmetry, center, and spread make sense only for quantitative variables.

(continued)

- **Don't use bars in every display—save them for histograms and bar charts.** In a bar chart, the bars indicate how many cases of a categorical variable are piled in each category. Bars in a histogram indicate the number of cases piled in each interval of a quantitative variable. In both bar charts and histograms, the bars represent counts of data values. Some people create other displays that use bars to represent individual data values. Beware: Such graphs are neither bar charts nor histograms. For example, a student was asked to make a histogram from data showing the number of juvenile bald eagles seen during each of the 13 weeks in the winter of 2003–2004 at a site in Rock Island, IL. Instead, he made this plot:

**FIGURE 4.15**

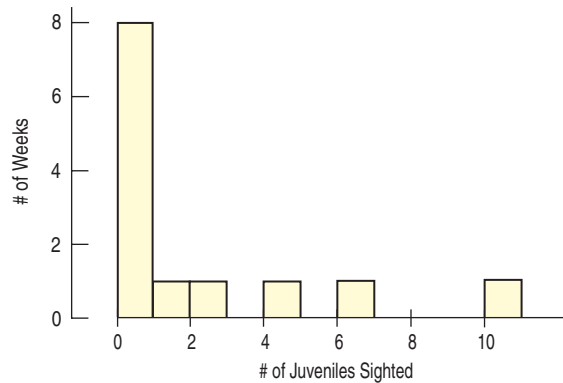
*This isn't a histogram or a bar chart. It's an ill-conceived graph that uses bars to represent individual data values (number of eagles sighted) week by week.*



Look carefully. That's not a histogram. A histogram shows *What* we've measured along the horizontal axis and counts of the associated *Who*'s represented as bar heights. This student has it backwards: He used bars to show counts of birds for each week.<sup>10</sup> We need counts of weeks. A correct histogram should have a tall bar at "0" to show there were many weeks when no eagles were seen, like this:

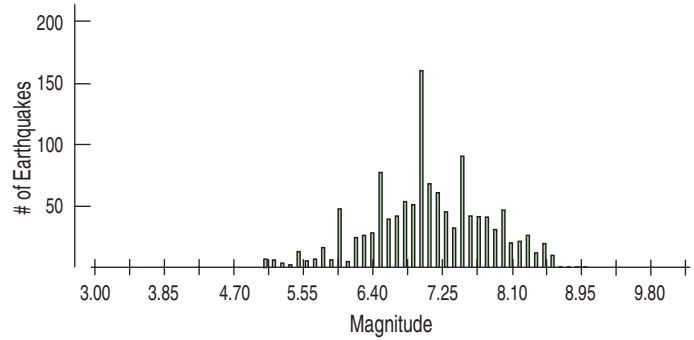
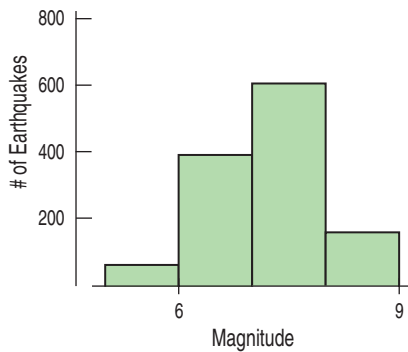
**FIGURE 4.16**

*A histogram of the eagle-sighting data shows the number of weeks in which different counts of eagles occurred. This display shows the distribution of juvenile-eagle sightings.*



- **Choose a bin width appropriate to the data.** Computer programs usually do a pretty good job of choosing histogram bin widths. Often there's an easy way to adjust the width, sometimes interactively. Here are the tsunami earthquakes with two (rather extreme) choices for the bin size:

<sup>10</sup> Edward Tufte, in his book *The Visual Display of Quantitative Information*, proposes that graphs should have a high data-to-ink ratio. That is, we shouldn't waste a lot of ink to display a single number when a dot would do the job.



The task of summarizing a quantitative variable is relatively simple, and there is a simple path to follow. However, you need to watch out for certain features of the data that make summarizing them with a number dangerous. Here's some advice:

- ▶ **Don't forget to do a reality check.** Don't let the computer or calculator do your thinking for you. Make sure the calculated summaries make sense. For example, does the mean look like it is in the center of the histogram? Think about the spread: An IQR of 50 mpg would clearly be wrong for gas mileage. And no measure of spread can be negative. The standard deviation can take the value 0, but only in the very unusual case that all the data values equal the same number. If you see an IQR or standard deviation equal to 0, it's probably a sign that something's wrong with the data.
- ▶ **Don't forget to sort the values before finding the median or percentiles.** It seems obvious, but when you work by hand, it's easy to forget to sort the data first before counting in to find medians, quartiles, or other percentiles. Don't report that the median of the five values 194, 5, 1, 17, and 893 is 1 just because 1 is the middle number.
- ▶ **Don't worry about small differences when using different methods.** Finding the 10th percentile or the lower quartile in a data set sounds easy enough. But it turns out that the definitions are not exactly clear. If you compare different statistics packages or calculators, you may find that they give slightly different answers for the same data. These differences, though, are unlikely to be important in interpreting the data, the quartiles, or the IQR, so don't let them worry you.

**Gold Card Customers—Regions National Banks**

Month	April 2007	May 2007
Average Zip Code	45,034.34	38,743.34

- ▶ **Don't compute numerical summaries of a categorical variable.** Neither the mean zip code nor the standard deviation of social security numbers is meaningful. If the variable is categorical, you should instead report summaries such as percentages of individuals in each category. It is easy to make this mistake when using technology to do the summaries for you. After all, the computer doesn't care what the numbers mean.
- ▶ **Don't report too many decimal places.** Statistical programs and calculators often report a ridiculous number of digits. A general rule for numerical summaries is to report one or two more digits than the number of digits in the data. For example, earlier we saw a dotplot of the Kentucky Derby race times. The mean and standard deviation of those times could be reported as:

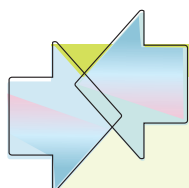
$$\bar{y} = 130.63401639344262 \text{ sec} \quad s = 13.66448201942662 \text{ sec}$$

But we knew the race times only to the nearest quarter second, so the extra digits are meaningless.

- ▶ **Don't round in the middle of a calculation.** Don't report too many decimal places, but it's best not to do any rounding until the end of your calculations. Even though you might report the mean of the earthquakes as 7.08, it's really 7.08339. Use the more precise number in your calculations if you're finding the standard deviation by hand—or be prepared to see small differences in your final result.

(continued)

- ▶ **Watch out for multiple modes.** The summaries of the Kentucky Derby times are meaningless for another reason. As we saw in the dotplot, the Derby was initially a longer race. It would make much more sense to report that the old 1.5 mile Derby had a mean time of 159.6 seconds, while the current Derby has a mean time of 124.6 seconds. If the distribution has multiple modes, consider separating the data into different groups and summarizing each group separately.
- ▶ **Beware of outliers.** The median and IQR are resistant to outliers, but the mean and standard deviation are not. To help spot outliers . . .
- ▶ **Don't forget to: Make a picture (make a picture, make a picture).** The sensitivity of the mean and standard deviation to outliers is one reason you should always make a picture of the data. Summarizing a variable with its mean and standard deviation when you have not looked at a histogram or dotplot to check for outliers or skewness invites disaster. You may find yourself drawing absurd or dangerously wrong conclusions about the data. And, of course, you should demand no less of others. Don't accept a mean and standard deviation blindly without some evidence that the variable they summarize is unimodal, symmetric, and free of outliers.



## CONNECTIONS

Distributions of quantitative variables, like those of categorical variables, show the possible values and their relative frequencies. A histogram shows the distribution of values in a quantitative variable with adjacent bars. Don't confuse histograms with bar charts, which display categorical variables. For categorical data, the mode is the category with the biggest count. For quantitative data, modes are peaks in the histogram.

The shape of the distribution of a quantitative variable is an important concept in most of the subsequent chapters. We will be especially interested in distributions that are unimodal and symmetric.

In addition to their shape, we summarize distributions with center and spread, usually pairing a measure of center with a measure of spread: median with IQR and mean with standard deviation. We favor the mean and standard deviation when the shape is unimodal and symmetric, but choose the median and IQR for skewed distributions or when there are outliers we can't otherwise set aside.

## WHAT HAVE WE LEARNED?



We've learned how to make a picture of quantitative data to help us see the story the data have to *Tell*.

- ▶ We can display the distribution of quantitative data with a *histogram*, a *stem-and-leaf* display, or a *dotplot*.
- ▶ We *Tell* what we see about the distribution by talking about *shape*, *center*, *spread*, and any *unusual features*.

We've learned how to summarize distributions of quantitative variables numerically.

- ▶ Measures of center for a distribution include the median and the mean.

We write the formula for the mean as  $\bar{y} = \frac{\sum y}{n}$ .

- ▶ Measures of spread include the range, IQR, and standard deviation.

The standard deviation is computed as  $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$ .

The median and IQR are not usually given as formulas.



- ▶ We'll report the median and IQR when the distribution is skewed. If it's symmetric, we'll summarize the distribution with the mean and standard deviation (and possibly the median and IQR as well). Always pair the median with the IQR and the mean with the standard deviation.

We've learned to *Think* about the type of variable we're summarizing.

- ▶ All the methods of this chapter assume that the data are quantitative.
- ▶ The **Quantitative Data Condition** serves as a check that the data are, in fact, quantitative. One good way to be sure is to know the measurement units. You'll want those as part of the *Think* step of your answers.

## Terms

Distribution	44. The distribution of a quantitative variable slices up all the possible values of the variable into equal-width bins and gives the number of values (or counts) falling into each bin.
Histogram (relative frequency histogram)	45. A histogram uses adjacent bars to show the distribution of a quantitative variable. Each bar represents the frequency (or relative frequency) of values falling in each bin.
Gap	45. A region of the distribution where there are no values.
Stem-and-leaf display	47. A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.
Dotplot	49. A dotplot graphs a dot for each case against a single axis.
Shape	49. To describe the shape of a distribution, look for <ul style="list-style-type: none"> <li>▶ single vs. multiple modes.</li> <li>▶ symmetry vs. skewness.</li> <li>▶ outliers and gaps.</li> </ul>
Center	52, 58. The place in the distribution of a variable that you'd point to if you wanted to attempt the impossible by summarizing the entire distribution with a single number. Measures of center include the mean and median.
Spread	54, 61. A numerical summary of how tightly the values are clustered around the center. Measures of spread include the IQR and standard deviation.
Mode	49. A hump or local high point in the shape of the distribution of a variable. The apparent location of modes can change as the scale of a histogram is changed.
Unimodal (Bimodal)	50. Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped. Distributions with two modes are called <b>bimodal</b> . Those with more than two are <b>multimodal</b> .
Uniform	50. A distribution that's roughly flat is said to be uniform.
Symmetric	50. A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
Tails	50. The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
Skewed	50. A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be <b>skewed left</b> when the longer tail stretches to the left, and <b>skewed right</b> when it goes to the right.
Outliers	51. Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or they may be just mistakes; there's no obvious way to tell. Don't delete outliers automatically—you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.
Median	52. The median is the middle value, with half of the data above and half below it. If $n$ is even, it is the average of the two middle values. It is usually paired with the IQR.
Range	54. The difference between the lowest and highest values in a data set. $Range = max - min$ .
Quartile	54. The lower quartile (Q1) is the value with a quarter of the data below it. The upper quartile (Q3) has three quarters of the data below it. The median and quartiles divide data into four parts with equal numbers of data values.

Interquartile range (IQR)	55. The IQR is the difference between the first and third quartiles. $IQR = Q3 - Q1$ . It is usually reported along with the median.
Percentile	55. The $i$ th percentile is the number that falls above $i\%$ of the data.
5-Number Summary	56. The 5-number summary of a distribution reports the minimum value, $Q1$ , the median, $Q3$ , and the maximum value.
Mean	58. The mean is found by summing all the data values and dividing by the count: $\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}.$ <p>It is usually paired with the standard deviation.</p>
Resistant	59. A calculated summary is said to be resistant if outliers have only a small effect on it.
Variance	61. The variance is the sum of squared deviations from the mean, divided by the count minus 1: $s^2 = \frac{\sum (y - \bar{y})^2}{n - 1}.$ <p>It is useful in calculations later in the book.</p>
Standard deviation	61. The standard deviation is the square root of the variance: $s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$ <p>It is usually reported along with the mean.</p>

## Skills

### THINK

- ▶ Be able to identify an appropriate display for any quantitative variable.
- ▶ Be able to guess the shape of the distribution of a variable by knowing something about the data.
- ▶ Be able to select a suitable measure of center and a suitable measure of spread for a variable based on information about its distribution.
- ▶ Know the basic properties of the median: The median divides the data into the half of the data values that are below the median and the half that are above.
- ▶ Know the basic properties of the mean: The mean is the point at which the histogram balances.
- ▶ Know that the standard deviation summarizes how spread out all the data are around the mean.
- ▶ Understand that the median and IQR resist the effects of outliers, while the mean and standard deviation do not.
- ▶ Understand that in a skewed distribution, the mean is pulled in the direction of the skewness (toward the longer tail) relative to the median.

### SHOW

- ▶ Know how to display the distribution of a quantitative variable with a stem-and-leaf display (drawn by hand for smaller data sets), a dotplot, or a histogram (made by computer for larger data sets).
- ▶ Know how to compute the mean and median of a set of data.
- ▶ Know how to compute the standard deviation and IQR of a set of data.

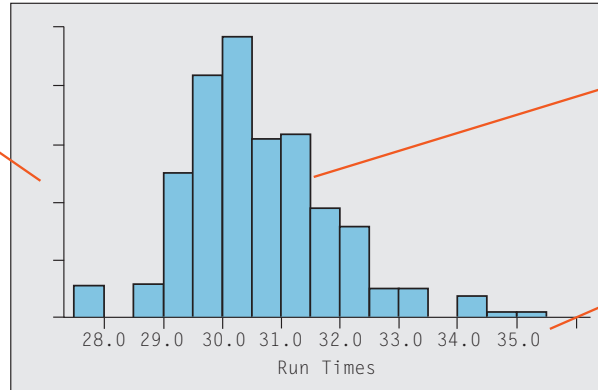
### TELL

- ▶ Be able to describe the distribution of a quantitative variable in terms of its shape, center, and spread.
- ▶ Be able to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Know how to describe summary measures in a sentence. In particular, know that the common measures of center and spread have the same units as the variable that they summarize, and should be described in those units.
- ▶ Be able to describe the distribution of a quantitative variable with a description of the shape of the distribution, a numerical measure of center, and a numerical measure of spread. Be sure to note any unusual features, such as outliers, too.

## DISPLAYING AND SUMMARIZING QUANTITATIVE VARIABLES ON THE COMPUTER

Almost any program that displays data can make a histogram, but some will do a better job of determining where the bars should start and how they should partition the span of the data.

The vertical scale may be counts or proportions. Sometimes it isn't clear which. But the shape of the histogram is the same either way.



Most packages choose the number of bars for you automatically. Often you can adjust that choice.

The axis should be clearly labeled so you can tell what "pile" each bar represents. You should be able to tell the lower and upper bounds of each bar.

Many statistics packages offer a prepackaged collection of summary measures. The result might look like this:

Variable: W eight  
 N = 234  
 Mean = 143.3                      Median = 139  
 St. Dev = 11.1                      IQR = 14

Alternatively, a package might make a table for several variables and summary measures:

**A S** **Case Study: Describing Distribution Shapes.** Who's safer in a crash—passengers or the driver? Investigate with your statistics package.

Variable	N	mean	median	stdev	IQR
Weight	234	143.3	139	11.1	14
Height	234	68.3	68.1	4.3	5
Score	234	86	88	9	5

It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

Packages often provide many more summary statistics than you need. Of course, some of these may not be appropriate when the data are skewed or have outliers. It is your responsibility to check a histogram or stem-and-leaf display and decide which summary statistics to use.

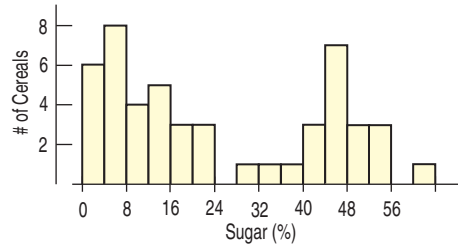
It is common for packages to report summary statistics to many decimal places of "accuracy." Of course, it is rare data that have such accuracy in the original measurements. The ability to calculate to six or seven digits beyond the decimal point doesn't mean that those digits have any meaning. Generally it's a good idea to round these values, allowing perhaps one more digit of precision than was given in the original data.

Displays and summaries of quantitative variables are among the simplest things you can do in most statistics packages.

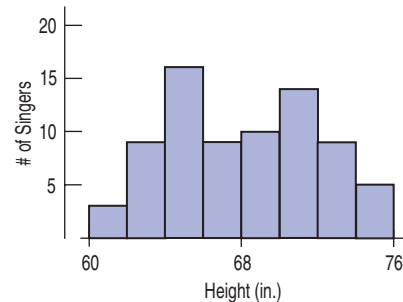
## EXERCISES

- Histogram.** Find a histogram that shows the distribution of a variable in a newspaper, a magazine, or the Internet.
  - Does the article identify the W's?
  - Discuss whether the display is appropriate.
  - Discuss what the display reveals about the variable and its distribution.
  - Does the article accurately describe and interpret the data? Explain.
- Not a histogram.** Find a graph other than a histogram that shows the distribution of a quantitative variable in a newspaper, a magazine, or the Internet.
  - Does the article identify the W's?
  - Discuss whether the display is appropriate for the data.
  - Discuss what the display reveals about the variable and its distribution.
  - Does the article accurately describe and interpret the data? Explain.
- In the news.** Find an article in a newspaper, a magazine, or the Internet that discusses an "average."
  - Does the article discuss the W's for the data?
  - What are the units of the variable?
  - Is the average used the median or the mean? How can you tell?
  - Is the choice of median or mean appropriate for the situation? Explain.
- In the news II.** Find an article in a newspaper, a magazine, or the Internet that discusses a measure of spread.
  - Does the article discuss the W's for the data?
  - What are the units of the variable?
  - Does the article use the range, IQR, or standard deviation?
  - Is the choice of measure of spread appropriate for the situation? Explain.
- Thinking about shape.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
  - The number of speeding tickets each student in the senior class of a college has ever had.
  - Players' scores (number of strokes) at the U.S. Open golf tournament in a given year.
  - Weights of female babies born in a particular hospital over the course of a year.
  - The length of the average hair on the heads of students in a large class.
- More shapes.** Would you expect distributions of these variables to be uniform, unimodal, or bimodal? Symmetric or skewed? Explain why.
  - Ages of people at a Little League game.
  - Number of siblings of people in your class.
  - Pulse rates of college-age males.
  - Number of times each face of a die shows in 100 tosses.

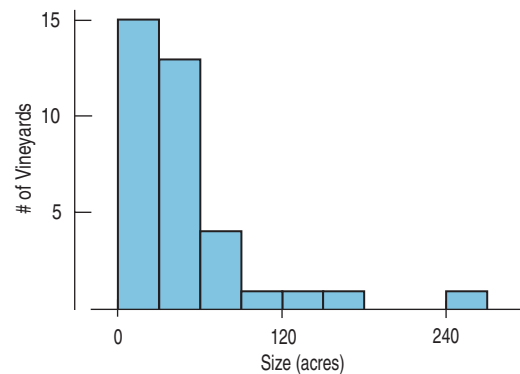
- T Sugar in cereals.** The histogram displays the sugar content (as a percent of weight) of 49 brands of breakfast cereals.



- Describe this distribution.
  - What do you think might account for this shape?
- T Singers.** The display shows the heights of some of the singers in a chorus, collected so that the singers could be positioned on stage with shorter ones in front and taller ones in back.

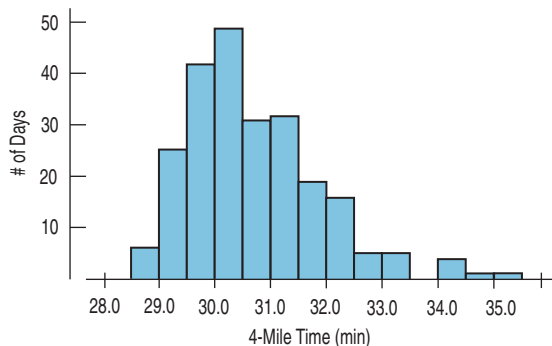


- Describe the distribution.
  - Can you account for the features you see here?
- T Vineyards.** The histogram shows the sizes (in acres) of 36 vineyards in the Finger Lakes region of New York.



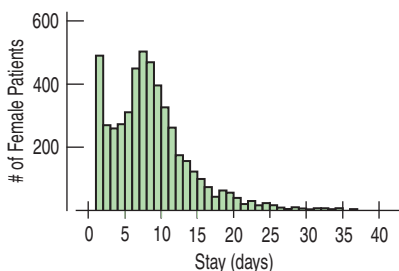
- Approximately what percentage of these vineyards are under 60 acres?
- Write a brief description of this distribution (shape, center, spread, unusual features).

10. **Run times.** One of the authors collected the times (in minutes) it took him to run 4 miles on various courses during a 10-year period. Here is a histogram of the times.



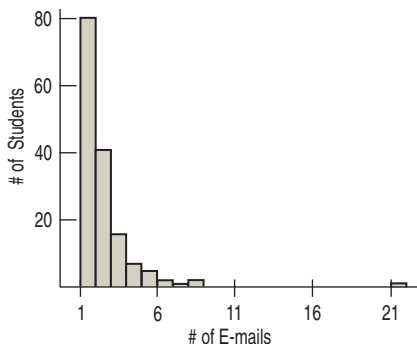
Describe the distribution and summarize the important features. What is it about running that might account for the shape you see?

11. **Heart attack stays.** The histogram shows the lengths of hospital stays (in days) for all the female patients admitted to hospitals in New York during one year with a primary diagnosis of acute myocardial infarction (heart attack).



- From the histogram, would you expect the mean or median to be larger? Explain.
- Write a few sentences describing this distribution (shape, center, spread, unusual features).
- Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T 12. **E-mails.** A university teacher saved every e-mail received from students in a large Introductory Statistics class during an entire term. He then counted, for each student who had sent him at least one e-mail, how many e-mails each student had sent.



- From the histogram, would you expect the mean or the median to be larger? Explain.
- Write a few sentences describing this distribution (shape, center, spread, unusual features).

- Which summary statistics would you choose to summarize the center and spread in these data? Why?

13. **Super Bowl points.** How many points do football teams score in the Super Bowl? Here are the total numbers of points scored by both teams in each of the first 42 Super Bowl games:

45, 47, 23, 30, 29, 27, 21, 31, 22, 38, 46, 37, 66, 50, 37, 47, 44, 47, 54, 56, 59, 52, 36, 65, 39, 61, 69, 43, 75, 44, 56, 55, 53, 39, 41, 37, 69, 61, 45, 31, 46, 31

- Find the median.
- Find the quartiles.
- Write a description based on the 5-number summary.

14. **Super Bowl wins.** In the Super Bowl, by how many points does the winning team outscore the losers? Here are the winning margins for the first 42 Super Bowl games:

25, 19, 9, 16, 3, 21, 7, 17, 10, 4, 18, 17, 4, 12, 17, 5, 10, 29, 22, 36, 19, 32, 4, 45, 1, 13, 35, 17, 23, 10, 14, 7, 15, 7, 27, 3, 27, 3, 3, 11, 12, 3

- Find the median.
- Find the quartiles.
- Write a description based on the 5-number summary.

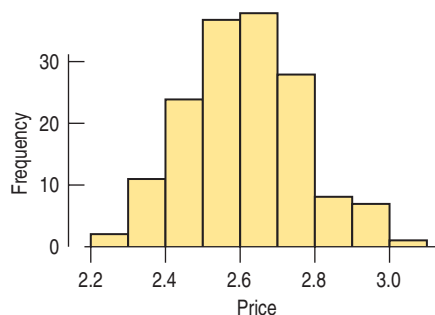
15. **Standard deviation I.** For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

Set 1	Set 2
a) 3, 5, 6, 7, 9	2, 4, 6, 8, 10
b) 10, 14, 15, 16, 20	10, 11, 15, 19, 20
c) 2, 6, 6, 9, 11, 14	82, 86, 86, 89, 91, 94

16. **Standard deviation II.** For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

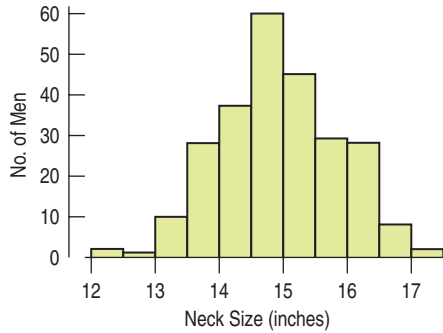
Set 1	Set 2
a) 4, 7, 7, 7, 10	4, 6, 7, 8, 10
b) 100, 140, 150, 160, 200	10, 50, 60, 70, 110
c) 10, 16, 18, 20, 22, 28	48, 56, 58, 60, 62, 70

- T 17. **Pizza prices.** The histogram shows the distribution of the prices of plain pizza slices (in \$) for 156 weeks in Dallas, TX.



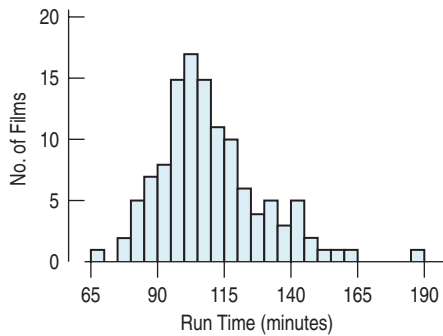
- Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T 18. Neck size.** The histogram shows the neck sizes (in inches) of 250 men recruited for a health study in Utah.

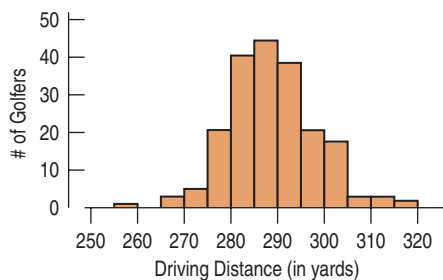


Which summary statistics would you choose to summarize the center and spread in these data? Why?

- T 19. Pizza prices again.** Look again at the histogram of the pizza prices in Exercise 17.
- Is the mean closer to \$2.40, \$2.60, or \$2.80? Why?
  - Is the standard deviation closer to \$0.15, \$0.50, or \$1.00? Explain.
- T 20. Neck sizes again.** Look again at the histogram of men's neck sizes in Exercise 18.
- Is the mean closer to 14, 15, or 16 inches? Why?
  - Is the standard deviation closer to 1 inch, 3 inches, or 5 inches? Explain.
- T 21. Movie lengths.** The histogram shows the running times in minutes of 122 feature films released in 2005.

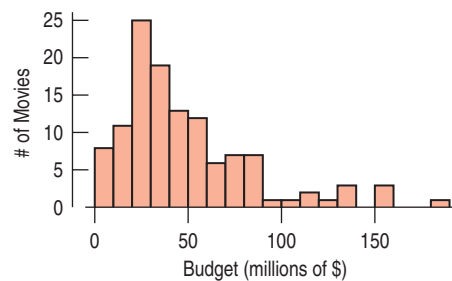


- You plan to see a movie this weekend. Based on these movies, how long do you expect a typical movie to run?
  - Would you be surprised to find that your movie ran for  $2\frac{1}{2}$  hours (150 minutes)?
  - Which would you expect to be higher: the mean or the median run time for all movies? Why?
- T 22. Golf drives.** The display shows the average drive distance (in yards) for 202 professional golfers on the men's PGA tour.



- Describe this distribution.
- Approximately what proportion of professional male golfers drive, on average, less than 280 yards?
- Estimate the mean by examining the histogram.
- Do you expect the mean to be smaller than, approximately equal to, or larger than the median? Why?

- 23. Movie lengths II.** Exercise 21 looked at the running times of movies released in 2005. The standard deviation of these running times is 19.6 minutes, and the quartiles are  $Q_1 = 97$  minutes and  $Q_3 = 119$  minutes.
- Write a sentence or two describing the spread in running times based on
    - the quartiles.
    - the standard deviation.
  - Do you have any concerns about using either of these descriptions of spread? Explain.
- 24. Golf drives II.** Exercise 22 looked at distances PGA golfers can hit the ball. The standard deviation of these average drive distances is 9.3 yards, and the quartiles are  $Q_1 = 282$  yards and  $Q_3 = 294$  yards.
- Write a sentence or two describing the spread in distances based on
    - the quartiles.
    - the standard deviation.
  - Do you have any concerns about using either of these descriptions of spread? Explain.
- 25. Mistake.** A clerk entering salary data into a company spreadsheet accidentally put an extra "0" in the boss's salary, listing it as \$2,000,000 instead of \$200,000. Explain how this error will affect these summary statistics for the company payroll:
- measures of center: median and mean.
  - measures of spread: range, IQR, and standard deviation.
- 26. Cold weather.** A meteorologist preparing a talk about global warming compiled a list of weekly low temperatures (in degrees Fahrenheit) he observed at his southern Florida home last year. The coldest temperature for any week was 36°F, but he inadvertently recorded the Celsius value of 2°. Assuming that he correctly listed all the other temperatures, explain how this error will affect these summary statistics:
- measures of center: mean and median.
  - measures of spread: range, IQR, and standard deviation.
- T 27. Movie budgets.** The histogram shows the budgets (in millions of dollars) of major release movies in 2005.



An industry publication reports that the average movie costs \$35 million to make, but a watchdog group con-

cerned with rising ticket prices says that the average cost is \$46.8 million. What statistic do you think each group is using? Explain.

28. **Sick days.** During contract negotiations, a company seeks to change the number of sick days employees may take, saying that the annual “average” is 7 days of absence per employee. The union negotiators counter that the “average” employee misses only 3 days of work each year. Explain how both sides might be correct, identifying the measure of center you think each side is using and why the difference might exist.
29. **Payroll.** A small warehouse employs a supervisor at \$1200 a week, an inventory manager at \$700 a week, six stock boys at \$400 a week, and four drivers at \$500 a week.
- Find the mean and median wage.
  - How many employees earn more than the mean wage?
  - Which measure of center best describes a typical wage at this company: the mean or the median?
  - Which measure of spread would best describe the payroll: the range, the IQR, or the standard deviation? Why?
30. **Singers.** The frequency table shows the heights (in inches) of 130 members of a choir.

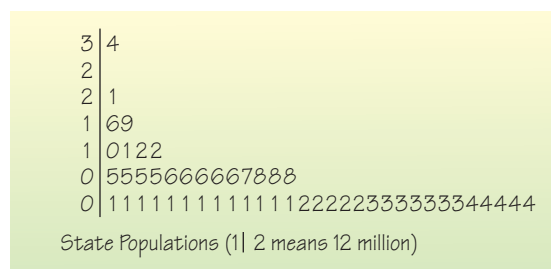
Height	Count	Height	Count
60	2	69	5
61	6	70	11
62	9	71	8
63	7	72	9
64	5	73	4
65	20	74	2
66	18	75	4
67	7	76	1
68	12		

- Find the median and IQR.
  - Find the mean and standard deviation.
  - Display these data with a histogram.
  - Write a few sentences describing the distribution.
31. **Gasoline.** In March 2006, 16 gas stations in Grand Junction, CO, posted these prices for a gallon of regular gasoline:

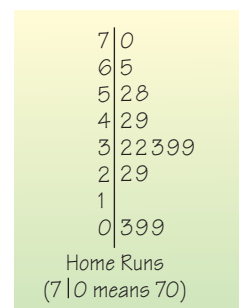
2.22	2.21	2.45	2.24
2.27	2.28	2.27	2.23
2.26	2.46	2.29	2.32
2.36	2.38	2.33	2.27

- Make a stem-and-leaf display of these gas prices. Use split stems; for example, use two 2.2 stems—one for prices between \$2.20 and \$2.24 and the other for prices from \$2.25 to \$2.29.
- Describe the shape, center, and spread of this distribution.
- What unusual feature do you see?

32. **The Great One.** During his 20 seasons in the NHL, Wayne Gretzky scored 50% more points than anyone who ever played professional hockey. He accomplished this amazing feat while playing in 280 fewer games than Gordie Howe, the previous record holder. Here are the number of games Gretzky played during each season: 79, 80, 80, 80, 74, 80, 80, 79, 64, 78, 73, 78, 74, 45, 81, 48, 80, 82, 82, 70
- Create a stem-and-leaf display for these data, using split stems.
  - Describe the shape of the distribution.
  - Describe the center and spread of this distribution.
  - What unusual feature do you see? What might explain this?
33. **States.** The stem-and-leaf display shows populations of the 50 states and Washington, DC, in millions of people, according to the 2000 census.



- What measures of center and spread are most appropriate?
  - Without doing any calculations, which must be larger: the median or the mean? Explain how you know.
  - From the stem-and-leaf display, find the median and the interquartile range.
  - Write a few sentences describing this distribution.
34. **Wayne Gretzky.** In Exercise 32, you examined the number of games played by hockey great Wayne Gretzky during his 20-year career in the NHL.
- Would you use the median or the mean to describe the center of this distribution? Why?
  - Find the median.
  - Without actually finding the mean, would you expect it to be higher or lower than the median? Explain.
35. **Home runs.** The stem-and-leaf display shows the number of home runs hit by Mark McGwire during the 1986–2001 seasons. Describe the distribution, mentioning its shape and any unusual features.



36. **Bird species.** The Cornell Lab of Ornithology holds an annual Christmas Bird Count ([www.birdsource.org](http://www.birdsource.org)), in which bird watchers at various locations around the country see how many different species of birds they can spot. Here are some of the counts reported from sites in Texas during the 1999 event:

228	178	186	162	206	166	163
183	181	206	177	175	167	162
160	160	157	156	153	153	152

- Create a stem-and-leaf display of these data.
- Write a brief description of the distribution. Be sure to discuss the overall shape as well as any unusual features.

37. **Hurricanes 2006.** The data below give the number of hurricanes classified as major hurricanes in the Atlantic Ocean each year from 1944 through 2006, as reported by NOAA ([www.nhc.noaa.gov](http://www.nhc.noaa.gov)):

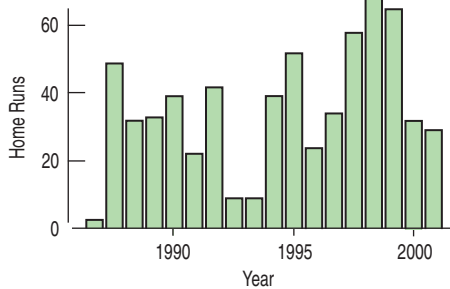
3, 2, 1, 2, 4, 3, 7, 2, 3, 3, 2, 5, 2, 2, 4, 2, 2, 6, 0, 2, 5, 1, 3, 1, 0, 3, 2, 1, 0, 1, 2, 3, 2, 1, 2, 2, 3, 1, 1, 1, 3, 0, 1, 3, 2, 1, 2, 1, 1, 0, 5, 6, 1, 3, 5, 3, 3, 2, 3, 6, 7, 2

- Create a dotplot of these data.
- Describe the distribution.

38. **Horsepower.** Create a stem-and-leaf display for these horsepowers of autos reviewed by *Consumer Reports* one year, and describe the distribution:

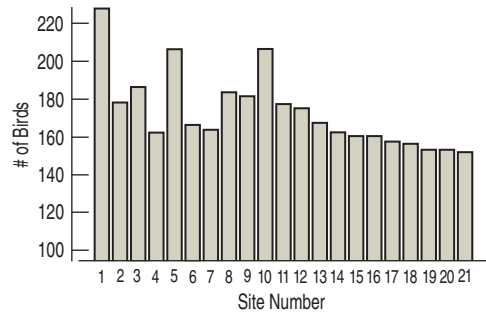
155	103	130	80	65
142	125	129	71	69
125	115	138	68	78
150	133	135	90	97
68	105	88	115	110
95	85	109	115	71
97	110	65	90	
75	120	80	70	

39. **Home runs again.** Students were asked to make a histogram of the number of home runs hit by Mark McGwire from 1986 to 2001 (see Exercise 35). One student submitted the following display:



- Comment on this graph.
- Create your own histogram of the data.

40. **Return of the birds.** Students were given the assignment to make a histogram of the data on bird counts reported in Exercise 36. One student submitted the following display:



- Comment on this graph.
- Create your own histogram of the data.

41. **Acid rain.** Two researchers measured the pH (a scale on which a value of 7 is neutral and values below 7 are acidic) of water collected from rain and snow over a 6-month period in Allegheny County, PA. Describe their data with a graph and a few sentences:

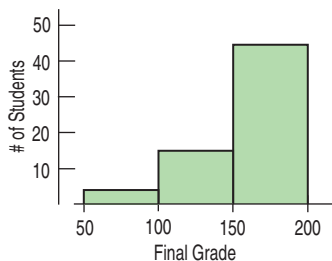
4.57	5.62	4.12	5.29	4.64	4.31	4.30	4.39	4.45
5.67	4.39	4.52	4.26	4.26	4.40	5.78	4.73	4.56
5.08	4.41	4.12	5.51	4.82	4.63	4.29	4.60	

42. **Marijuana 2003.** In 2003 the Council of Europe published a report entitled *The European School Survey Project on Alcohol and Other Drugs* ([www.espad.org](http://www.espad.org)). Among other issues, the survey investigated the percentages of 16-year-olds who had used marijuana. Shown here are the results for 20 European countries. Create an appropriate graph of these data, and describe the distribution.

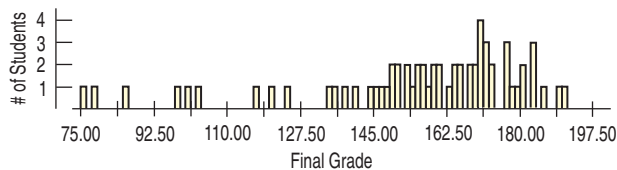
Country	Percentage	Country	Percentage
Austria	21%	Italy	27%
Belgium	32%	Latvia	16%
Bulgaria	21%	Lithuania	13%
Croatia	22%	Malta	10%
Cyprus	4%	Netherlands	28%
Czech Republic	44%	Norway	9%
Denmark	23%	Poland	18%
Estonia	23%	Portugal	15%
Faroe Islands	9%	Romania	3%
Finland	11%	Russia	22%
France	22%	Slovak Republic	27%
Germany	27%	Slovenia	28%
Greece	6%	Sweden	7%
Greenland	27%	Switzerland	40%
Hungary	16%	Turkey	4%
Iceland	13%	Ukraine	21%
Ireland	39%	United Kingdom	38%
Isle of Man	39%		



43. **Final grades.** A professor (of something other than Statistics!) distributed the following histogram to show the distribution of grades on his 200-point final exam. Comment on the display.

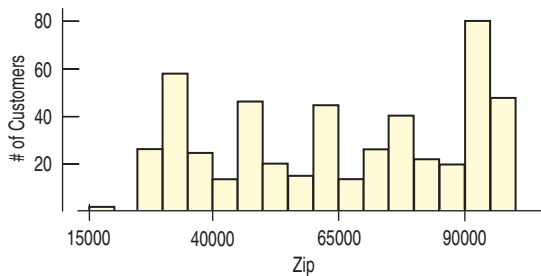


44. **Final grades revisited.** After receiving many complaints about his final-grade histogram from students currently taking a Statistics course, the professor from Exercise 43 distributed the following revised histogram:



- a) Comment on this display.  
b) Describe the distribution of grades.

45. **Zip codes.** Holes-R-U's, an Internet company that sells piercing jewelry, keeps transaction records on its sales. At a recent sales meeting, one of the staff presented a histogram of the zip codes of the last 500 customers, so that the staff might understand where sales are coming from. Comment on the usefulness and appropriateness of the display.



46. **Zip codes revisited.** Here are some summary statistics to go with the histogram of the zip codes of 500 customers from the Holes-R-U's Internet Jewelry Salon that we saw in Exercise 45:

Count	500
Mean	64,970.0
StdDev	23,523.0
Median	64,871
IQR	44,183
Q1	46,050
Q3	90,233

What can these statistics tell you about the company's sales?

47. **Math scores 2005.** The National Center for Education Statistics (<http://nces.ed.gov/nationsreportcard/>) reported 2005 average mathematics achievement scores for eighth graders in all 50 states:

State	Score	State	Score
Alabama	225	Montana	241
Alaska	236	Nebraska	238
Arizona	230	Nevada	230
Arkansas	236	New Hampshire	246
California	230	New Jersey	244
Colorado	239	New Mexico	224
Connecticut	242	New York	238
Delaware	240	North Carolina	241
Florida	239	North Dakota	243
Georgia	234	Ohio	242
Hawaii	230	Oklahoma	234
Idaho	242	Oregon	238
Illinois	233	Pennsylvania	241
Indiana	240	Rhode Island	233
Iowa	240	South Carolina	238
Kansas	246	South Dakota	242
Kentucky	231	Tennessee	232
Louisiana	230	Texas	242
Maine	241	Utah	239
Maryland	238	Vermont	244
Massachusetts	247	Virginia	240
Michigan	238	Washington	242
Minnesota	246	West Virginia	231
Mississippi	227	Wisconsin	241
Missouri	235	Wyoming	243

- a) Find the median, the IQR, the mean, and the standard deviation of these state averages.  
b) Which summary statistics would you report for these data? Why?  
c) Write a brief summary of the performance of eighth graders nationwide.

48. **Boomtowns.** In 2006, *Inc.* magazine ([www.inc.com](http://www.inc.com)) listed its choice of "boomtowns" in the United States—larger cities that are growing rapidly. Here is the magazine's top 20, along with their job growth percentages:

City	1-Year Job Growth (%)
Las Vegas, NV	7.5
Fort Lauderdale, FL	4.2
Orlando, FL	4.5
West Palm Beach-Boca Raton, FL	3.4
San Bernadino-Riverside, CA	1.9
Phoenix, AZ	4.4
Northern Virginia, VA	3.1
Washington, DC-Arlington-Alexandria, VA	3.2
Tampa-St. Petersburg, FL	2.6
Camden-Burlington counties, NJ	2.6

(continued)

City	1-Year Job Growth (%)
Jacksonville, FL	2.6
Charlotte, NC	3.3
Raleigh-Cary, NC	2.8
Richmond, VA	2.9
Salt Lake City, UT	3.3
Putnam-Rockland-Westchester counties, New York	2.3
Santa Ana-Anaheim-Irvine, CA	1.7
Miami-Miami Beach, FL	2.2
Sacramento, CA	1.5
San Diego, CA	1.4

Massachusetts	458.5	Oklahoma	614.2
Michigan	482.0	Oregon	418.4
Minnesota	527.7	Pennsylvania	386.8
Mississippi	558.5	Rhode Island	454.6
Missouri	550.5	South Carolina	578.6
Montana	544.4	South Dakota	564.4
Nebraska	470.1	Tennessee	552.5
Nevada	367.9	Texas	532.7
New Hampshire	544.4	Utah	460.6
New Jersey	488.2	Vermont	545.5
New Mexico	508.8	Virginia	526.9
New York	293.4	Washington	423.6
North Carolina	505.0	West Virginia	426.7
North Dakota	553.7	Wisconsin	449.8
Ohio	451.1	Wyoming	615.0

- Make a suitable display of the growth rates.
- Summarize the typical growth rate among these cities with a median and mean. Why do they differ?
- Given what you know about the distribution, which of the measures in b) does the better job of summarizing the growth rates? Why?
- Summarize the spread of the growth rate distribution with a standard deviation and with an IQR.
- Given what you know about the distribution, which of the measures in d) does the better job of summarizing the growth rates? Why?
- Suppose we subtract from each of the preceding growth rates the predicted U.S. average growth rate of 1.20%, so that we can look at how much these growth rates exceed the U.S. rate. How would this change the values of the summary statistics you calculated above? (*Hint:* You need not recompute any of the summary statistics from scratch.)
- If we were to omit Las Vegas from the data, how would you expect the mean, median, standard deviation, and IQR to change? Explain your expectations for each.
- Write a brief report about all of these growth rates.

- T 49. Gasoline usage 2004.** The California Energy Commission ([www.energy.ca.gov/gasoline/](http://www.energy.ca.gov/gasoline/)) collects data on the amount of gasoline sold in each state. The following data show the per capita (gallons used per person) consumption in the year 2004. Using appropriate graphical displays and summary statistics, write a report on the gasoline use by state in the year 2004.

State	Gallons per Capita	State	Gallons per Capita
Alabama	529.4	Hawaii	358.7
Alaska	461.7	Idaho	454.8
Arizona	381.9	Illinois	408.3
Arkansas	512.0	Indiana	491.7
California	414.4	Iowa	555.1
Colorado	435.7	Kansas	511.8
Connecticut	435.7	Kentucky	526.6
Delaware	541.6	Louisiana	507.8
Florida	496.0	Maine	576.3
Georgia	537.1	Maryland	447.5

- T 50. Prisons 2005.** A report from the U.S. Department of Justice ([www.ojp.usdoj.gov/bjs/](http://www.ojp.usdoj.gov/bjs/)) reported the percent changes in federal prison populations in 21 northeastern and midwestern states during 2005. Using appropriate graphical displays and summary statistics, write a report on the changes in prison populations.

State	Percent Change	State	Percent Change
Connecticut	-0.3	Iowa	2.5
Maine	0.0	Kansas	1.1
Massachusetts	5.5	Michigan	1.4
New Hampshire	3.3	Minnesota	6.0
New Jersey	2.2	Missouri	-0.8
New York	-1.6	Nebraska	7.9
Pennsylvania	3.5	North Dakota	4.4
Rhode Island	6.5	Ohio	2.3
Vermont	5.6	South Dakota	11.9
Illinois	2.0	Wisconsin	-1.0
Indiana	1.9		



## **JUST CHECKING**

### **Answers**

(Thoughts will vary.)

- 1.** Roughly symmetric, slightly skewed to the right. Center around 3 miles? Few over 10 miles.
- 2.** Bimodal. Center between 1 and 2 hours? Many people watch no football; others watch most of one or more games. Probably only a few values over 5 hours.
- 3.** Strongly skewed to the right, with almost everyone at \$0; a few small prizes, with the winner an outlier.
- 4.** Fairly symmetric, somewhat uniform, perhaps slightly skewed to the right. Center in the 40s? Few ages below 25 or above 70.
- 5.** Uniform, symmetric. Center near 5. Roughly equal counts for each digit 0–9.
- 6.** Incomes are probably skewed to the right and not symmetric, making the median the more appropriate measure of center. The mean will be influenced by the high end of family incomes and not reflect the “typical” family income as well as the median would. It will give the impression that the typical income is higher than it is.
- 7.** An IQR of 30 mpg would mean that only 50% of the cars get gas mileages in an interval 30 mpg wide. Fuel economy doesn’t vary that much. 3 mpg is reasonable. It seems plausible that 50% of the cars will be within about 3 mpg of each other. An IQR of 0.3 mpg would mean that the gas mileage of half the cars varies little from the estimate. It’s unlikely that cars, drivers, and driving conditions are that consistent.
- 8.** We’d prefer a standard deviation of 2 months. Making a consistent product is important for quality. Customers want to be able to count on the MP3 player lasting somewhere close to 5 years, and a standard deviation of 2 years would mean that life-spans were highly variable.

# Understanding and Comparing Distributions



<b>WHO</b>	Days during 1989
<b>WHAT</b>	Average daily wind speed (mph), Average barometric pressure (mb), Average daily temperature (deg Celsius)
<b>WHEN</b>	1989
<b>WHERE</b>	Hopkins Forest, in Western Massachusetts
<b>WHY</b>	Long-term observations to study ecology and climate

The Hopkins Memorial Forest is a 2500-acre reserve in Massachusetts, New York, and Vermont managed by the Williams College Center for Environmental Studies (CES). As part of their mission, CES monitors forest resources and conditions over the long term. They post daily measurements at their Web site.<sup>1</sup> You can go there, download, and analyze data for any range of days. We'll focus for now on 1989. As we'll see, some interesting things happened that year.

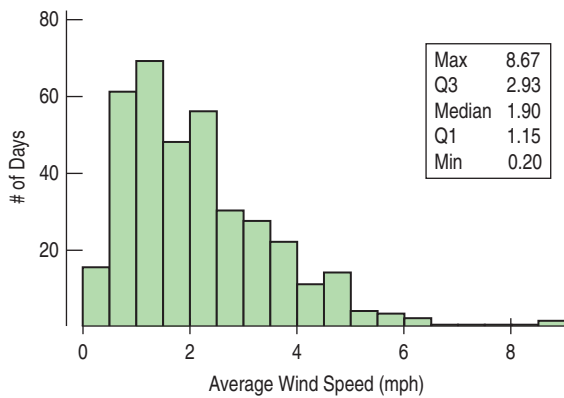
One of the variables measured in the forest is wind speed. Three remote anemometers generate far too much data to report, so, as summaries, you'll find the minimum, maximum, and average wind speed (in mph) for each day.

Wind is caused as air flows from areas of high pressure to areas of low pressure. Centers of low pressure often accompany storms, so both high winds and low pressure are associated with some of the fiercest storms. Wind speeds can vary greatly during a day and from day to day, but if we step back a bit farther, we can see patterns. By modeling these patterns, we can understand things about *Average Wind Speed* that we may not have known.

In Chapter 3 we looked at the association between two categorical variables using contingency tables and displays. Here we'll explore different ways of examining the relationship between two variables when one is quantitative, and the other is categorical and indicates groups to compare. We are given wind speed averages for each day of 1989. But we can collect the days together into different size groups and compare the wind speeds among them. If we consider *Time* as a categorical variable in this way, we'll gain enormous flexibility for our analysis and for our understanding. We'll discover new insights as we change the granularity of the grouping variable—from viewing the whole year's data at one glance, to comparing seasons, to looking for patterns across months, and, finally, to looking at the data day by day.

<sup>1</sup> [www.williams.edu/CES/hopkins.htm](http://www.williams.edu/CES/hopkins.htm)

## The Big Picture

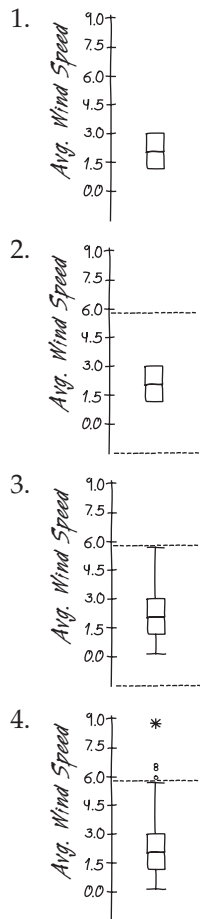


Let's start with the "big picture." Here's a histogram and 5-number summary of the *Average Wind Speed* for every day in 1989. Because of the skewness, we'll report the median and IQR. We can see that the distribution of *Average Wind Speed* is unimodal and skewed to the right. Median daily wind speed is about 1.90 mph, and on half of the days, the average wind speed is between 1.15 and 2.93 mph. We also see a rather windy 8.67-mph day. Was that unusually windy or just the windiest day of the year? To answer that, we'll need to work with the summaries a bit more.

FIGURE 5.1

A histogram of daily Average Wind Speed for 1989. It is unimodal and skewed to the right, with a possible high outlier.

## Boxplots and 5-Number Summaries



Once we have a 5-number summary of a (quantitative) variable, we can display that information in a **boxplot**. To make a boxplot of the average wind speeds, follow these steps:

1. Draw a single vertical axis spanning the extent of the data.<sup>2</sup> Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box. The box can have any width that looks OK.<sup>3</sup>
2. To help us construct the boxplot, we erect "fences" around the main part of the data. We place the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile. For the wind speed data, we compute

$$\text{Upper fence} = Q3 + 1.5 \text{ IQR} = 2.93 + 1.5 \times 1.78 = 5.60 \text{ mph}$$

and

$$\text{Lower fence} = Q1 - 1.5 \text{ IQR} = 1.15 - 1.5 \times 1.78 = -1.52 \text{ mph}$$

The fences are just for construction and are not part of the display. We show them here with dotted lines for illustration. You should never include them in your boxplot.

3. We use the fences to grow "whiskers." Draw lines from the ends of the box up and down to the most extreme data values found within the fences. If a data value falls outside one of the fences, we do *not* connect it with a whisker.
4. Finally, we add the **outliers** by displaying any data values beyond the fences with special symbols. (We often use a different symbol for "far outliers"—data values farther than 3 IQRs from the quartiles.)

What does a boxplot show? The center of a boxplot is (remarkably enough) a box that shows the middle half of the data, between the quartiles. The height of the box is equal to the IQR. If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If the median is not centered, the distribution is skewed. The whiskers show skewness as well if they are not roughly the same length. Any outliers are displayed individually, both to keep them out of the way for judging skewness and to encourage you to give them special attention. They may be mistakes, or they may be the most interesting cases in your data.

**A S** **Boxplots.** Watch a boxplot under construction.

### TI-*n*spire

**Boxplots and dotplots.** Drag data points around to explore what a boxplot shows (and doesn't).

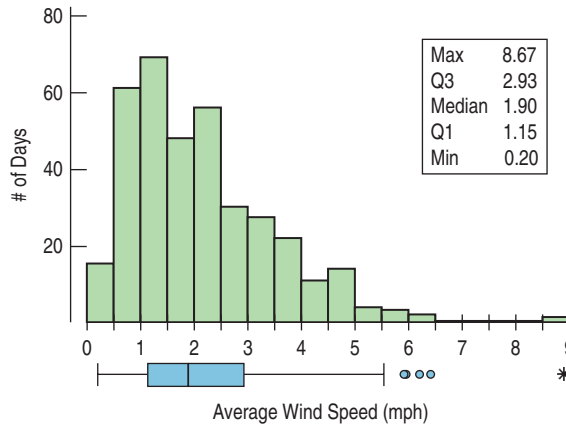
<sup>2</sup>The axis could also run horizontally.

<sup>3</sup>Some computer programs draw wider boxes for larger data sets. That can be useful when comparing groups.

The prominent statistician John W. Tukey, the originator of the boxplot, was asked by one of the authors why the outlier nomination rule cut at 1.5 IQRs beyond each quartile. He answered that the reason was that 1 IQR would be too small and 2 IQRs would be too large. That works for us.

**AS** **Activity: Playing with Summaries.** See how different summary measures behave as you place and drag values, and see how sensitive some statistics are to individual data values.

For the Hopkins Forest data, the central box contains each day whose *Average Wind Speed* is between 1.15 and 2.93 miles per hour (see Figure 5.2). From the shape of the box, it looks like the central part of the distribution of wind speeds is roughly symmetric, but the longer upper whisker indicates that the distribution stretches out at the upper end. We also see a few very windy days. Boxplots are particularly good at pointing out outliers. These extraordinarily windy days may deserve more attention. We'll give them that extra attention shortly.

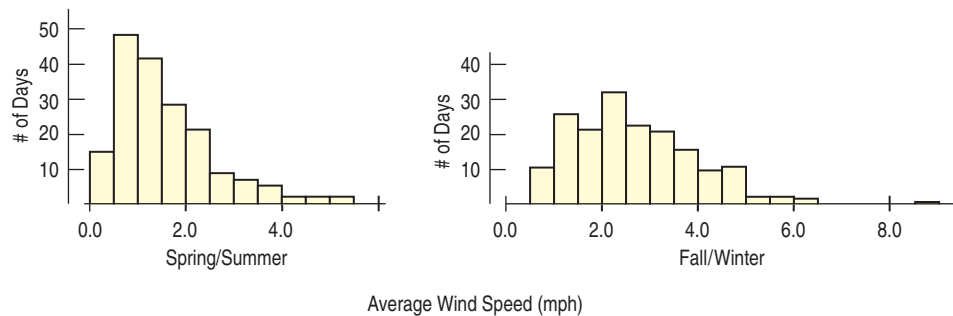


**FIGURE 5.2**  
By turning the boxplot and putting it on the same scale as the histogram, we can compare both displays of the daily wind speeds and see how each represents the distribution.

## Comparing Groups with Histograms

**TI-*n*spire**  
**Histograms and boxplots.** See that the shape of a distribution is not always evident in a boxplot.

It is almost always more interesting to compare groups. Is it windier in the winter or the summer? Are any months particularly windy? Are weekends a special problem? Let's split the year into two groups: April through September (Spring/Summer) and October through March (Fall/Winter). To compare the groups, we create two histograms, being careful to use the same scale. Here are displays of the average daily wind speed for Spring/Summer (on the left) and Fall/Winter (on the right):



**FIGURE 5.3**  
Histograms of Average Wind Speed for days in Spring/Summer (left) and Fall/Winter (right) show very different patterns.

The shapes, centers, and spreads of these two distributions are strikingly different. During spring and summer (histogram on the left), the distribution is skewed to the right. A typical day during these warmer months has an average wind speed of only 1 to 2 mph, and few have average speeds above 3 mph. In the colder months (histogram on the right), however, the shape is less strongly skewed and more spread out. The typical wind speed is higher, and days with average wind speeds above 3 mph are not unusual. There are several noticeable high values.

Summaries for Average Wind Speed by Season				
Group	Mean	StdDev	Median	IQR
Fall/Winter	2.71	1.36	2.47	1.87
Spring/Summer	1.56	1.01	1.34	1.32

**FOR EXAMPLE**

**Comparing groups with stem-and-leaf displays**

In 2004 the infant death rate in the United States was 6.8 deaths per 1000 live births. The Kaiser Family Foundation collected data from all 50 states and the District of Columbia, allowing us to look at different regions of the country. Since there are only 51 data values, a back-to-back stem-and-leaf plot is an effective display. Here's one comparing infant death rates in the Northeast and Midwest to those in the South and West. In this display the stems run down the middle of the plot, with the leaves for the two regions to the left or right. Be careful when you read the values on the left: 4 | 11 | means a rate of 11.4 deaths per 1000 live birth for one of the southern or western states.

**Infant Death Rates (by state) 2004**

South and West		North and Midwest
	4	11
	3	10
	0	9
0 4 1 6 9 5 8	8	10
0 5 0 3	7	5 8 0 7 4 1
4 1 0 4 9 1 1 6 4	6	3 1 5 4 4
6 3 6 2	5	8 4 0 6
	4	8 8 9 7
	3	

(4 | 11 | means 11.4 deaths per 1000 live births)

**Question:** How do infant death rates compare for these regions?

In general, infant death rates were generally higher for states in the South and West than in the Northeast and Midwest. The distribution for the northeastern and midwestern states is roughly uniform, varying from a low of 4.8 to a high of 8.1 deaths per 1000 live births. Ten southern and western states had higher infant death rates than any in the Northeast or Midwest, with one state over 11. Rates varied more widely in the South and West, where the distribution is skewed to the right and possibly bimodal. We should investigate further to see which states represent the cluster of high death rates.

## Comparing Groups with Boxplots

**AS**

**Video: Can Diet Prolong**

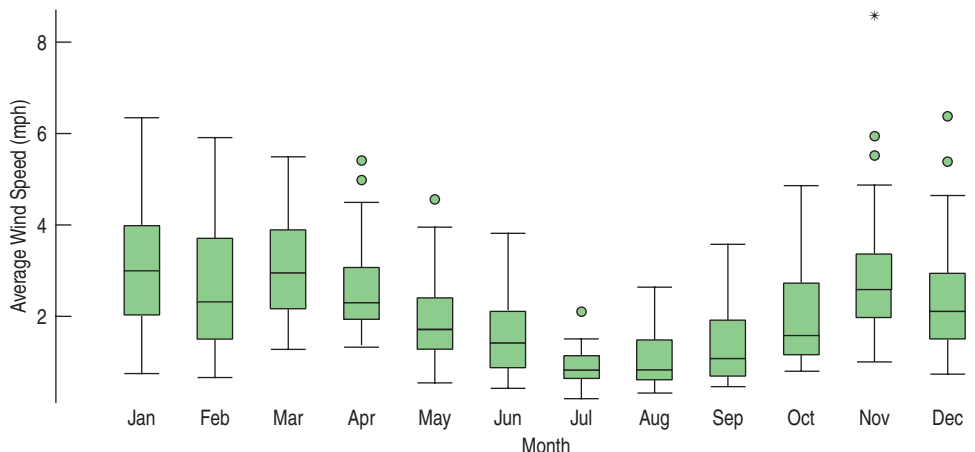
**Life?** Here's a subject that's been in the news: Can you live longer by eating less? (Or would it just seem longer?) Look at the data in subsequent activities, and you'll find that you can learn a lot by comparing two groups with boxplots.

Are some months windier than others? Even residents may not have a good idea of which parts of the year are the most windy. (Do you know for your hometown?) We're not interested just in the centers, but also in the spreads. Are wind speeds equally variable from month to month, or do some months show more variation?

Earlier, we compared histograms of the wind speeds for two halves of the year. To look for seasonal trends, though, we'll group the daily observations by month. Histograms or stem-and-leaf displays are a fine way to look at one distribution or two. But it would be hard to see patterns by comparing 12 histograms. Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information. So we often plot them side by side for groups or categories we wish to compare.

By placing boxplots side by side, we can easily see which groups have higher medians, which have the greater IQRs, where the central 50% of the data is located in each group, and which have the greater overall range. And, when the boxes are in an order, we can get a general idea of patterns in both the centers and the spreads. Equally important, we can see past any outliers in making these comparisons because they've been displayed separately.

Here are boxplots of the *Average Daily Wind Speed* by month:



**FIGURE 5.4**

Boxplots of the average daily wind speed for each month show seasonal patterns in both the centers and spreads.

Here we see that wind speeds tend to decrease in the summer. The months in which the winds are both strongest and most variable are November through March. And there was one remarkably windy day in November.

When we looked at a boxplot of wind speeds for the entire year, there were only 5 outliers. Now, when we group the days by *Month*, the boxplots display more days as outliers and call out one in November as a far outlier. The boxplots show different outliers than before because some days that seemed ordinary when placed against the entire year's data looked like outliers for the month that they're in. That windy day in July certainly wouldn't stand out in November or December, but for July, it was remarkable.

## FOR EXAMPLE

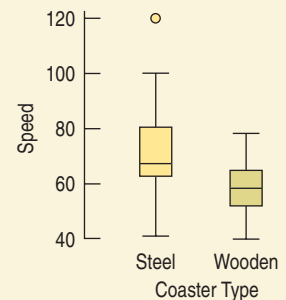
### Comparing distributions

Roller coasters<sup>4</sup> are a thrill ride in many amusement parks worldwide. And thrill seekers want a coaster that goes fast. There are two main types of roller coasters: those with wooden tracks and those with steel tracks. Do they typically run at different speeds? Here are boxplots:

**Question:** Compare the speeds of wood and steel roller coasters.



Overall, wooden-track roller coasters are slower than steel-track coasters. In fact, the fastest half of the steel coasters are faster than three quarters of the wooden coasters. Although the IQRs of the two groups are similar, the range of speeds among steel coasters is larger than the range for wooden coasters. The distribution of speeds of wooden coasters appears to be roughly symmetric, but the speeds of the steel coasters are skewed to the right, and there is a high outlier at 120 mph. We should look into why that steel coaster is so fast.



## STEP-BY-STEP EXAMPLE

### Comparing Groups

Of course, we can compare groups even when they are not in any particular order. Most scientific studies compare two or more groups. It is almost always a good idea to start an analysis of data from such studies by comparing boxplots for the groups. Here's an example:

For her class project, a student compared the efficiency of various coffee containers. For her study, she decided to try 4 different containers and to test each of them 8 different times. Each time, she heated water to 180°F, poured it into a container, and sealed it. (We'll learn the details of how to set up experiments in Chapter 13.) After 30 minutes, she measured the temperature again and recorded the difference in temperature. Because these are temperature differences, smaller differences mean that the liquid stayed hot—just what we would want in a coffee mug.

**Question:** What can we say about the effectiveness of these four mugs?

<sup>4</sup> See the Roller Coaster Data Base at [www.rcdb.com](http://www.rcdb.com).





**Plan** State what you want to find out.

**Variables** Identify the *variables* and report the *W*'s.

Be sure to check the appropriate condition.

I want to compare the effectiveness of the different mugs in maintaining temperature. I have 8 measurements of *Temperature Change* for each of the mugs.

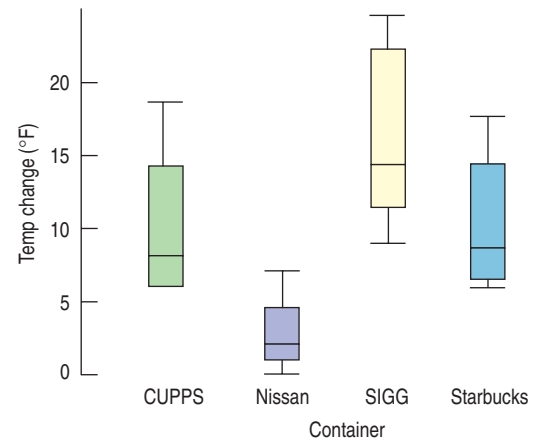
✓ **Quantitative Data Condition:** The *Temperature Changes* are quantitative, with units of °F. Boxplots are appropriate displays for comparing the groups. Numerical summaries of each group are appropriate as well.



**Mechanics** Report the 5-number summaries of the four groups. Including the IQR is a good idea as well.

Make a picture. Because we want to compare the distributions for four groups, boxplots are an appropriate choice.

	Min	Q1	Median	Q3	Max	IQR
CUPPS	6°F	6	8.25	14.25	18.50	8.25
Nissan	0	1	2	4.50	7	3.50
SIGG	9	11.50	14.25	21.75	24.50	10.25
Starbucks	6	6.50	8.50	14.25	17.50	7.75



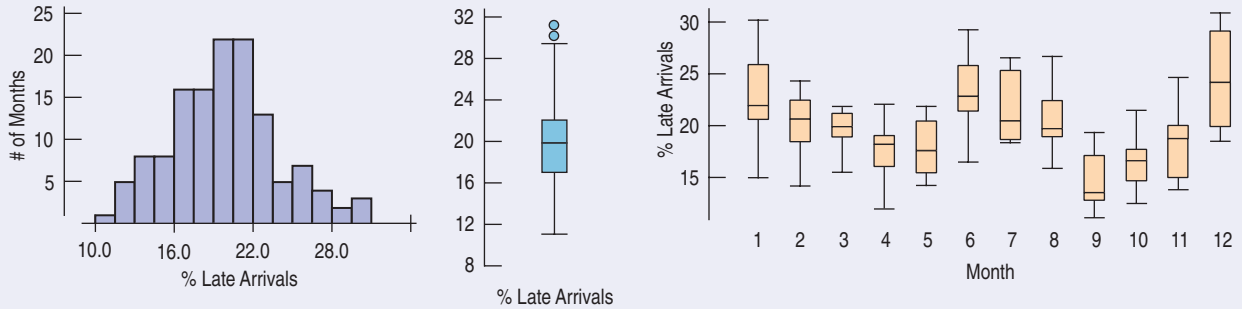
**Conclusion** Interpret what the boxplots and summaries say about the ability of these mugs to retain heat. Compare the shapes, centers, and spreads, and note any outliers.

The individual distributions of temperature changes are all slightly skewed to the high end. The Nissan cup does the best job of keeping liquids hot, with a median loss of only 2°F, and the SIGG cup does the worst, typically losing 14°F. The difference is large enough to be important: A coffee drinker would be likely to notice a 14° drop in temperature. And the mugs are clearly different: 75% of the Nissan tests showed less heat loss than any of the other mugs in the study. The IQR of results for the Nissan cup is also the smallest of these test cups, indicating that it is a consistent performer.



### JUST CHECKING

The Bureau of Transportation Statistics of the U.S. Department of Transportation collects and publishes statistics on airline travel ([www.transtats.bts.gov](http://www.transtats.bts.gov)). Here are three displays of the % of flights arriving late each month from 1995 through 2005:



1. Describe what the histogram says about late arrivals.
2. What does the boxplot of late arrivals suggest that you can't see in the histogram?
3. Describe the patterns shown in the boxplots by month. At what time of year are flights least likely to be late? Can you suggest reasons for this pattern?

#### T1 Tips

### Comparing groups with boxplots

In the last chapter we looked at the performances of fourth-grade students on an agility test. Now let's make comparative boxplots for the boys' scores and the girls' scores:

*Boys:* 22, 17, 18, 29, 22, 22, 23, 24, 23, 17, 21

*Girls:* 25, 20, 12, 19, 28, 24, 22, 21, 25, 26, 25, 16, 27, 22

Enter these data in **L1** (*Boys*) and **L2** (*Girls*).

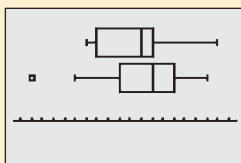
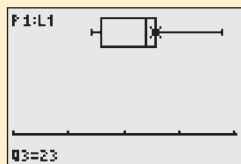
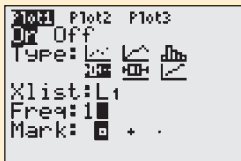
Set up **STATPLOT**'s **Plot1** to make a boxplot of the boys' data:

- Turn the plot **On**;
- Choose the first boxplot icon (you want your plot to indicate outliers);
- Specify **Xlist:L1** and **Freq:1**, and select the **Mark** you want the calculator to use for displaying any outliers.

Use **ZoomStat** to display the boxplot for *Boys*. You can now **TRACE** to see the statistics in the five-number summary. Try it!

As you did for the boys, set up **Plot2** to display the girls' data. This time when you use **ZoomStat** with both plots turned on, the display shows the parallel boxplots. See the outlier?

This is a great opportunity to practice your "Tell" skills. How do these fourth graders compare in terms of agility?



## Outliers

When we looked at boxplots for the *Average Wind Speed by Month*, we noticed that several days stood out as possible outliers and that one very windy day in November seemed truly remarkable. What should we do with such outliers?

Cases that stand out from the rest of the data almost always deserve our attention. An outlier is a value that doesn't fit with the rest of the data, but exactly how different it should be to be treated specially is a judgment call. Boxplots provide a rule of thumb to highlight these unusual points, but that rule doesn't tell you what to do with them.

So, what *should* we do with outliers? The first thing to do is to try to understand them in the context of the data. A good place to start is with a histogram. Histograms show us more detail about a distribution than a boxplot can, so they give us a better idea of how the outlier fits (or doesn't fit) in with the rest of the data.

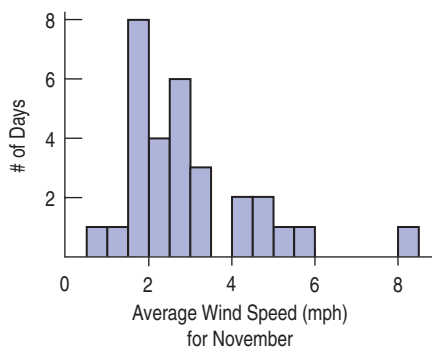
A histogram of the *Average Wind Speed* in November shows a slightly skewed main body of data and that very windy day clearly set apart from the other days. When considering whether a case is an outlier, we often look at the gap between that case and the rest of the data. A large gap suggests that the case really is quite different. But a case that just happens to be the largest or smallest value at the end of a possibly stretched-out tail may be best thought of as just . . . the largest or smallest value. After all, *some* case has to be the largest or smallest.

Some outliers are simply unbelievable. If a class survey includes a student who claims to be 170 inches tall (about 14 feet, or 4.3 meters), you can be pretty sure that's an error.

Once you've identified likely outliers, you should always investigate them. Some outliers are just errors. A decimal point may have been misplaced, digits transposed, or digits repeated or omitted. The units may be wrong. (Was that outlying height reported in centimeters rather than in inches [170 cm = 65 in.]?) Or a number may just have been transcribed incorrectly, perhaps copying an adjacent value on the original data sheet. If you can identify the correct value, then you should certainly fix it. One important reason to look into outliers is to correct errors in your data.

Many outliers are not wrong; they're just different. Such cases often repay the effort to understand them. You can learn more from the extraordinary cases than from summaries of the overall data set.

What about that windy November day? Was it really that windy, or could there have been a problem with the anemometers? A quick Internet search for weather on November 21, 1989, finds that there was a severe storm:



**FIGURE 5.5**

*The Average Wind Speed in November is slightly skewed with a high outlier.*



### **WIND, SNOW, COLD GIVE N.E. A TASTE OF WINTER**

*Published on November 22, 1989*

*Author: Andrew Dabilis, Globe Staff*

An intense storm roared like the Montreal Express through New England yesterday, bringing frigid winds of up to 55 m.p.h., 2 feet of snow in some parts of Vermont and a preview of winter after weeks of mild weather. Residents throughout the region awoke yesterday to an icy vortex that lifted an airplane off the runway in Newark and made driving dangerous in New England because of rapidly shifting winds that seemed to come from all directions.

When we have outliers, we need to decide what to *Tell* about the data. If we can correct an error, we'll just summarize the corrected data (and note the correction). But if we see no way to correct an outlying value, or if we confirm that it is correct, our best path is to report summaries and analyses with *and* without the outlier. In this way a reader can judge for him- or herself what influence the outlier has and decide what to think about the data.

There are two things we should *never* do with outliers. The first is to silently leave an outlier in place and proceed as if nothing were unusual. Analyses of data with outliers are very likely to be influenced by those outliers—sometimes to a large and misleading degree. The other is to drop an outlier from the analysis without comment just because it's unusual. If you want to exclude an outlier, you must discuss your decision and, to the extent you can, justify your decision.

**AS** **Case Study: Are passengers or drivers safer in a crash?** Practice the skills of this chapter by comparing these two groups.

## FOR EXAMPLE

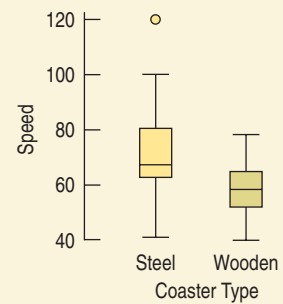
### Checking out the outliers

**Recap:** We've looked at the speeds of roller coasters and found a difference between steel- and wooden-track coasters. We also noticed an extraordinary value.

**Question:** The fastest coaster in this collection turns out to be the "Top Thrill Dragster" at Cedar Point amusement park. What might make this roller coaster unusual? You'll have to do some research, but that's often what happens with outliers.

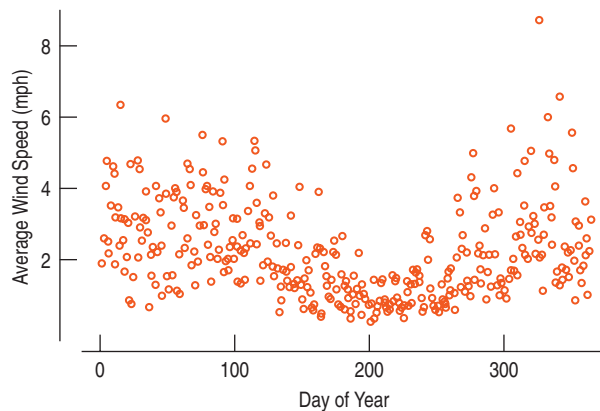
The Top Thrill Dragster is easy to find in an Internet search. We learn that it is a "hydraulic launch" coaster. That is, it doesn't get its remarkable speed just from gravity, but rather from a kick-start by a hydraulic piston. That could make it different from the other roller coasters.

(You might also discover that it is no longer the fastest roller coaster in the world.)



## Timeplots: Order, Please!

The Hopkins Forest wind speeds are reported as daily averages. Previously, we grouped the days into months or seasons, but we could look at the wind speed values day by day. Whenever we have data measured over time, it is a good idea to look for patterns by plotting the data in time order. Here are the daily average wind speeds plotted over time:



**FIGURE 5.6**

A timeplot of Average Wind Speed shows the overall pattern and changes in variation.

A display of values against time is sometimes called a **timeplot**. This timeplot reflects the pattern that we saw when we plotted the wind speeds by month. But without the arbitrary divisions between months, we can see a calm period during the summer, starting around day 200 (the middle of July), when the wind is relatively mild and doesn't vary greatly from day to day. We can also see that the wind becomes both more variable and stronger during the early and late parts of the year.

## Looking into the Future

It is always tempting to try to extend what we see in a timeplot into the future. Sometimes that makes sense. Most likely, the Hopkins Forest climate follows regular seasonal patterns. It's probably safe to predict a less windy June next year and a windier November. But we certainly wouldn't predict another storm on November 21.

Other patterns are riskier to extend into the future. If a stock has been rising, will it continue to go up? No stock has ever increased in value indefinitely, and no stock analyst has consistently been able to forecast when a stock's value will turn around. Stock prices, unemployment rates, and other economic, social, or psychological concepts are much harder to predict than physical quantities. The path a ball will follow when thrown from a certain height at a given speed and direction is well understood. The path interest rates will take is much less clear. Unless we have strong (nonstatistical) reasons for doing otherwise, we should resist the temptation to think that any trend we see will continue, even into the near future.

Statistical models often tempt those who use them to think beyond the data. We'll pay close attention later in this book to understanding when, how, and how much we can justify doing that.

## Re-expressing Data: A First Look

### RE-EXPRESSING TO IMPROVE SYMMETRY

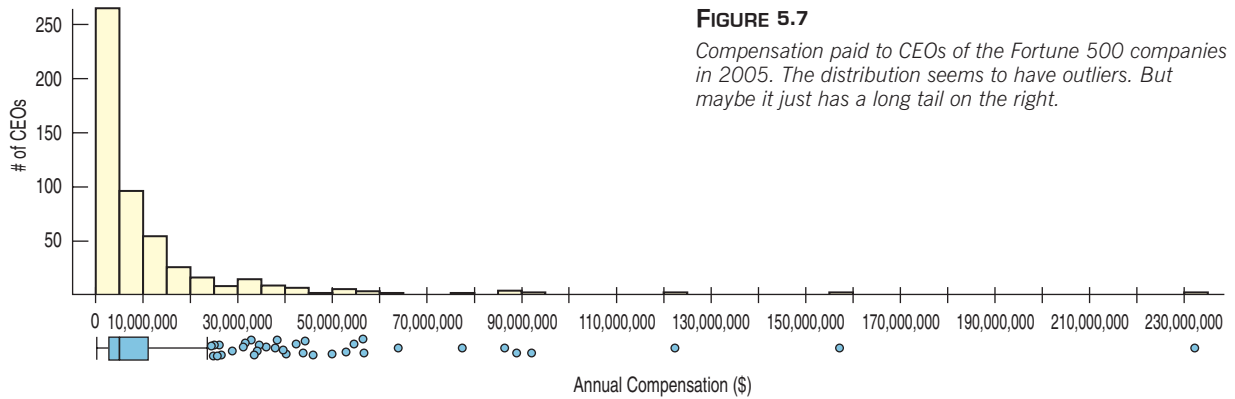
When the data are skewed, it can be hard to summarize them simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of the stretched-out tail. How can we say anything useful about such data? The secret is to *re-express* the data by applying a simple function to each value.

Many relationships and "laws" in the sciences and social sciences include functions such as logarithms, square roots, and reciprocals. Similar relationships often show up in data. Here's a simple example:

In 1980 large companies' chief executive officers (CEOs) made, on average, about 42 times what workers earned. In the next two decades, CEO compensation soared when compared to the average worker. By 2000 that multiple had jumped<sup>5</sup>

<sup>5</sup> Sources: United for a Fair Economy, *Business Week* annual CEO pay surveys, Bureau of Labor Statistics, "Average Weekly Earnings of Production Workers, Total Private Sector." Series ID: EEU00500004.

to 525. What does the distribution of the compensation of Fortune 500 companies' CEOs look like? Here's a histogram and boxplot for 2005 compensation:



**FIGURE 5.7**  
 Compensation paid to CEOs of the Fortune 500 companies in 2005. The distribution seems to have outliers. But maybe it just has a long tail on the right.

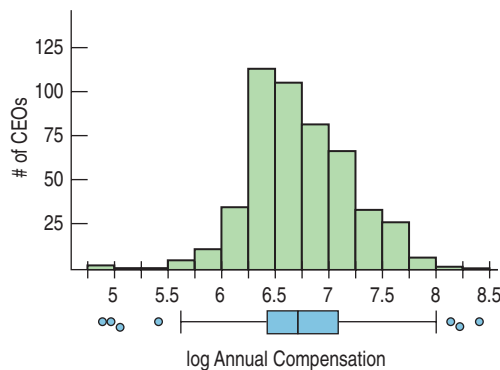
We have 500 CEOs and about 48 possible histogram bins, most of which are empty—but don't miss the tiny bars straggling out to the right. The boxplot indicates that some CEOs received extraordinarily high compensations, while the majority received relatively "little." But look at the values of the bins. The first bin, with about half the CEOs, covers incomes from \$0 to \$5,000,000. Imagine receiving a salary survey with these categories:

- What is your income?
- a) \$0 to \$5,000,000
  - b) \$5,000,001 to \$10,000,000
  - c) \$10,000,001 to \$15,000,000
  - d) More than \$15,000,000

The reason that the histogram seems to leave so much of the area blank is that the salaries are spread all along the axis from about \$15,000,000 to \$240,000,000. After \$50,000,000 there are so few for each bin that it's very hard to see the tiny bars. What we *can* see from this histogram and boxplot is that this distribution is highly skewed to the right.

It can be hard to decide what we mean by the "center" of a skewed distribution, so it's hard to pick a typical value to summarize the distribution. What would you say was a typical CEO total compensation? The mean value is \$10,307,000, while the median is "only" \$4,700,000. Each tells us something different about the data.

One approach is to **re-express, or transform, the data by applying a simple function to make the skewed distribution more symmetric.** For example, we could take the square root or logarithm of each compensation value. Taking logs works pretty well for the CEO compensations, as you can see:

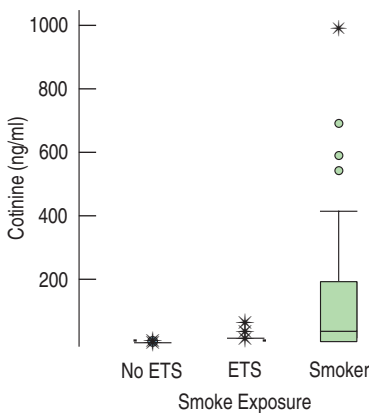


**FIGURE 5.8**  
 The logarithms of 2005 CEO compensations are much more nearly symmetric.

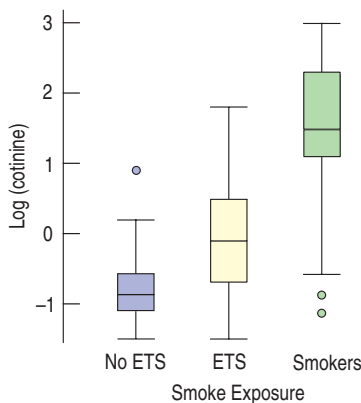
The histogram of the logs of the total CEO compensations is much more nearly symmetric, so we can see that a typical log compensation is between 6, which corresponds to \$1,000,000, and 7, corresponding to \$10,000,000. And it's easier to talk about a typical value for the logs. The mean log compensation is 6.73, while the median is 6.67. (That's \$5,370,317 and \$4,677,351, respectively.) Notice that nearly all the values are between 6.0 and 8.0—in other words, between \$1,000,000 and \$100,000,000 a year, but who's counting?

Against the background of a generally symmetric main body of data, it's easier to decide whether the largest compensations are outliers. In fact, the three most highly compensated CEOs are identified as outliers by the boxplot rule of thumb even after this re-expression. It's perhaps impressive to be an outlier CEO in annual compensation. It's even more impressive to be an outlier in the log scale!

**Dealing with logarithms** You have probably learned about logs in math courses and seen them in psychology or science classes. In this book, we use them only for making data behave better. Base 10 logs are the easiest to understand, but natural logs are often used as well. (Either one is fine.) You can think of base 10 logs as roughly one less than the number of digits you need to write the number. So 100, which is the smallest number to require 3 digits, has a  $\log_{10}$  of 2. And 1000 has a  $\log_{10}$  of 3. The  $\log_{10}$  of 500 is between 2 and 3, but you'd need a calculator to find that it's approximately 2.7. All salaries of "six figures" have  $\log_{10}$  between 5 and 6. Logs are incredibly useful for making skewed data more symmetric. But don't worry—nobody does logs without technology and neither should you. Often, remaking a histogram or other display of the data is as easy as pushing another button.



**FIGURE 5.9**  
Cotinine levels (nanograms per milliliter) for three groups with different exposures to tobacco smoke. Can you compare the ETS (exposed to smoke) and No-ETS groups?



**FIGURE 5.10**  
Blood cotinine levels after taking logs. What a difference a log makes!

## RE-EXPRESSION TO EQUALIZE SPREAD ACROSS GROUPS

Researchers measured the concentration (nanograms per milliliter) of cotinine in the blood of three groups of people: nonsmokers who have not been exposed to smoke, nonsmokers who have been exposed to smoke (ETS), and smokers. Cotinine is left in the blood when the body metabolizes nicotine, so this measure gives a direct measurement of the effect of passive smoke exposure. The boxplots of the cotinine levels of the three groups tell us that the smokers have higher cotinine levels, but if we want to compare the levels of the passive smokers to those of the nonsmokers, we're in trouble, because on this scale, the cotinine levels for both nonsmoking groups are too low to be seen.

Re-expressing can help alleviate the problem of comparing groups that have very different spreads. For measurements like the cotinine data, whose values can't be negative and whose distributions are skewed to the high end, a good first guess at a re-expression is the logarithm.

After taking logs, we can compare the groups and see that the nonsmokers exposed to environmental smoke (the ETS group) do show increased levels of (log) cotinine, although not the high levels found in the blood of smokers.

Notice that the same re-expression has also improved the symmetry of the cotinine distribution for smokers and pulled in most of the apparent outliers in all of the groups. It is not unusual for a re-expression that improves one aspect of data to improve others as well. We'll talk about other ways to re-express data as the need arises throughout the book. We'll explore some common re-expressions more thoroughly in Chapter 10.

## WHAT CAN GO WRONG?

► **Avoid inconsistent scales.** Parts of displays should be mutually consistent—no fair changing scales in the middle or plotting two variables on different scales but on the same display. When comparing two groups, be sure to compare them on the same scale.

► **Label clearly.** Variables should be identified clearly and axes labeled so a reader knows what the plot displays.

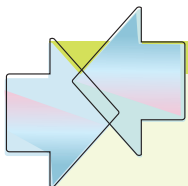
Here's a remarkable example of a plot gone wrong. It illustrated a news story about rising college costs. It uses time-plots, but it gives a misleading impression. First think about the story you're being told by this display. Then try to figure out what has gone wrong.

What's wrong? Just about everything.

- The horizontal scales are inconsistent. Both lines show trends over time, but exactly for what years? The tuition sequence starts in 1965, but rankings are graphed from 1989. Plotting them on the same (invisible) scale makes it seem that they're for the same years.
- The vertical axis isn't labeled. That hides the fact that it's inconsistent. Does it graph dollars (of tuition) or ranking (of Cornell University)?

This display violates three of the rules. And it's even worse than that: It violates a rule that we didn't even bother to mention.

- The two inconsistent scales for the vertical axis don't point in the same direction! The line for Cornell's rank shows that it has "plummeted" from 15th place to 6th place in academic rank. Most of us think that's an *improvement*, but that's not the message of this graph.
- **Beware of outliers.** If the data have outliers and you can correct them, you should do so. If they are clearly wrong or impossible, you should remove them and report on them. Otherwise, consider summarizing the data both with and without the outliers.



## CONNECTIONS

We discussed the value of summarizing a distribution with shape, center, and spread in Chapter 4, and we developed several ways to measure these attributes. Now we've seen the value of comparing distributions for different groups and of looking at patterns in a quantitative variable measured over time. Although it can be interesting to summarize a single variable for a single group, it is almost always more interesting to compare groups and look for patterns across several groups and over time. We'll continue to make comparisons like these throughout the rest of our work.



## WHAT HAVE WE LEARNED?



- ▶ We've learned the value of comparing groups and looking for patterns among groups and over time.
- ▶ We've seen that boxplots are very effective for comparing groups graphically. When we compare groups, we discuss their shape, center, and spreads, and any unusual features.
- ▶ We've experienced the value of identifying and investigating outliers. And we've seen that when we group data in different ways, it can allow different cases to emerge as possible outliers.
- ▶ We've graphed data that have been measured over time against a time axis and looked for long-term trends.

### Terms

Boxplot	81. A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups and for displaying outliers.
Outlier	81, 87. Any point more than 1.5 IQR from either end of the box in a boxplot is nominated as an outlier.
Far Outlier	81. If a point is more than 3.0 IQR from either end of the box in a boxplot, it is nominated as a <i>far outlier</i> .
Comparing distributions	82. When comparing the distributions of several groups using histograms or stem-and-leaf displays, consider their: <ul style="list-style-type: none"> <li>▶ Shape</li> <li>▶ Center</li> <li>▶ Spread</li> </ul>
Comparing boxplots	83. When comparing groups with boxplots: <ul style="list-style-type: none"> <li>▶ Compare the shapes. Do the boxes look symmetric or skewed? Are there differences between groups?</li> <li>▶ Compare the medians. Which group has the higher center? Is there any pattern to the medians?</li> <li>▶ Compare the IQRs. Which group is more spread out? Is there any pattern to how the IQRs change?</li> <li>▶ Using the IQRs as a background measure of variation, do the medians seem to be different, or do they just vary much as you'd expect from the overall variation?</li> <li>▶ Check for possible outliers. Identify them if you can and discuss why they might be unusual. Of course, correct them if you find that they are errors.</li> </ul>
Timeplot	88. A timeplot displays data that change over time. Often, successive values are connected with lines to show trends more clearly. Sometimes a smooth curve is added to the plot to help show long-term patterns and trends.

### Skills

#### THINK

- ▶ Be able to select a suitable display for comparing groups. Understand that histograms show distributions well, but are difficult to use when comparing more than two or three groups. Boxplots are more effective for comparing several groups, in part because they show much less information about the distribution of each group.
- ▶ Understand that how you group data can affect what kinds of patterns and relationships you are likely to see. Know how to select groupings to show the information that is important for your analysis.
- ▶ Be aware of the effects of skewness and outliers on measures of center and spread. Know how to select appropriate measures for comparing groups based on their displayed distributions.
- ▶ Understand that outliers can emerge at different groupings of data and that, whatever their source, they deserve special attention.
- ▶ Recognize when it is appropriate to make a timeplot.

SHOW

- ▶ Know how to make side-by-side histograms on comparable scales to compare the distributions of two groups.
- ▶ Know how to make side-by-side boxplots to compare the distributions of two or more groups.
- ▶ Know how to describe differences among groups in terms of patterns and changes in their center, spread, shape, and unusual values.
- ▶ Know how to make a timeplot of data that have been measured over time.

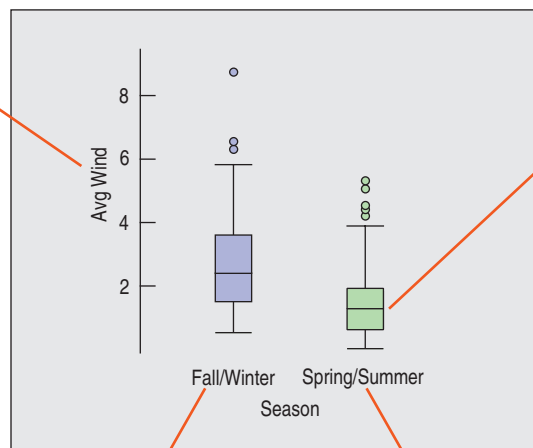
TELL

- ▶ Know how to compare the distributions of two or more groups by comparing their shapes, centers, and spreads. Be prepared to explain your choice of measures of center and spread for comparing the groups.
- ▶ Be able to describe trends and patterns in the centers and spreads of groups—especially if there is a natural order to the groups, such as a time order.
- ▶ Be prepared to discuss patterns in a timeplot in terms of both the general trend of the data and the changes in how spread out the pattern is.
- ▶ Be cautious about assuming that trends over time will continue into the future.
- ▶ Be able to describe the distribution of a quantitative variable in terms of its shape, center, and spread.
- ▶ Be able to describe any anomalies or extraordinary features revealed by the display of a variable.
- ▶ Know how to compare the distributions of two or more groups by comparing their shapes, centers, and spreads.
- ▶ Know how to describe patterns over time shown in a timeplot.
- ▶ Be able to discuss any outliers in the data, noting how they deviate from the overall pattern of the data.

## COMPARING DISTRIBUTIONS ON THE COMPUTER

Most programs for displaying and analyzing data can display plots to compare the distributions of different groups. Typically these are boxplots displayed side-by-side.

Side-by-side boxplots should be on the same y-axis scale so they can be compared.



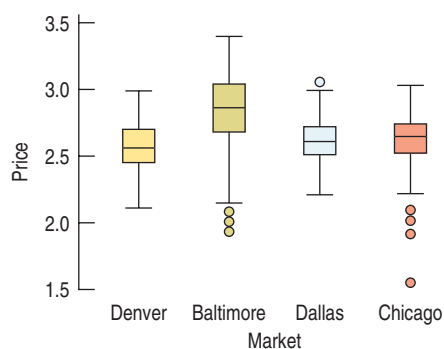
Some programs offer a graphical way to assess how much the medians differ by drawing a band around the median or by “notching” the boxes.

Boxes are typically labeled with a group name. Often they are placed in alphabetical order by group name—not the most useful order.

## EXERCISES

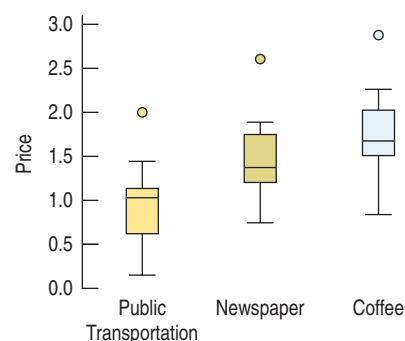
- In the news.** Find an article in a newspaper, magazine, or the Internet that compares two or more groups of data.
  - Does the article discuss the W's?
  - Is the chosen display appropriate? Explain.
  - Discuss what the display reveals about the groups.
  - Does the article accurately describe and interpret the data? Explain.
- In the news.** Find an article in a newspaper, magazine, or the Internet that shows a time plot.
  - Does the article discuss the W's?
  - Is the timeplot appropriate for the data? Explain.
  - Discuss what the timeplot reveals about the variable.
  - Does the article accurately describe and interpret the data? Explain.
- Time on the Internet.** Find data on the Internet (or elsewhere) that give results recorded over time. Make an appropriate display and discuss what it shows.
- Groups on the Internet.** Find data on the Internet (or elsewhere) for two or more groups. Make appropriate displays to compare the groups, and interpret what you find.

- T** 5. **Pizza prices.** A company that sells frozen pizza to stores in four markets in the United States (Denver, Baltimore, Dallas, and Chicago) wants to examine the prices that the stores charge for pizza slices. Here are boxplots comparing data from a sample of stores in each market:



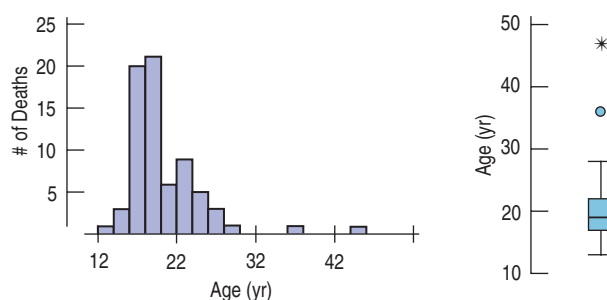
- Do prices appear to be the same in the four markets? Explain.
- Does the presence of any outliers affect your overall conclusions about prices in the four markets?

- T** 6. **Costs.** To help travelers know what to expect, researchers collected the prices of commodities in 16 cities throughout the world. Here are boxplots comparing the prices of a ride on public transportation, a newspaper, and a cup of coffee in the 16 cities (prices are all in \$US).



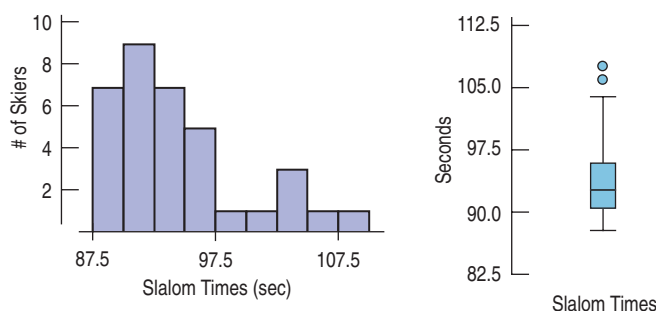
- On average, which commodity is the most expensive?
- Is a newspaper always more expensive than a ride on public transportation? Explain.
- Does the presence of outliers affect your conclusions in a) or b)?

- T** 7. **Still rockin'.** Crowd Management Strategies monitors accidents at rock concerts. In their database, they list the names and other variables of victims whose deaths were attributed to "crowd crush" at rock concerts. Here are the histogram and boxplot of the victims' ages for data from 1999 to 2000:



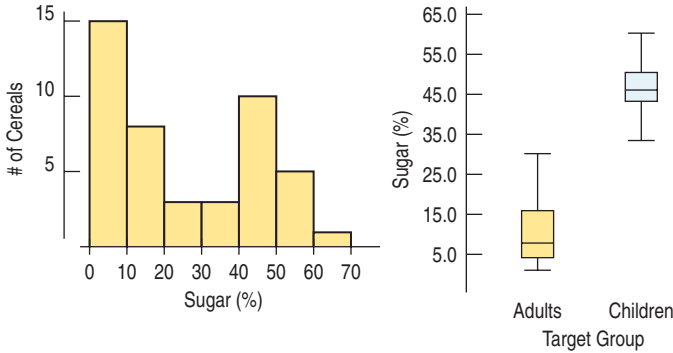
- What features of the distribution can you see in both the histogram and the boxplot?
- What features of the distribution can you see in the histogram that you could not see in the boxplot?
- What summary statistic would you choose to summarize the center of this distribution? Why?
- What summary statistic would you choose to summarize the spread of this distribution? Why?

- T** 8. **Slalom times.** The Men's Combined skiing event consists of a downhill and a slalom. Here are two displays of the slalom times in the Men's Combined at the 2006 Winter Olympics:



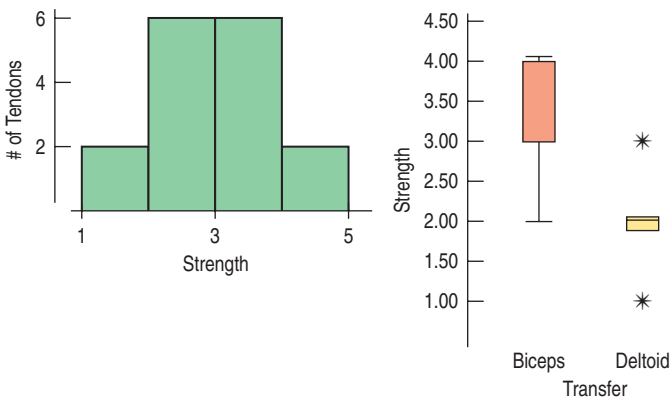
- a) What features of the distribution can you see in both the histogram and the boxplot?
- b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
- c) What summary statistic would you choose to summarize the center of this distribution? Why?
- d) What summary statistic would you choose to summarize the spread of this distribution? Why?

**T 9. Cereals.** Sugar is a major ingredient in many breakfast cereals. The histogram displays the sugar content as a percentage of weight for 49 brands of cereal. The boxplot compares sugar content for adult and children’s cereals.



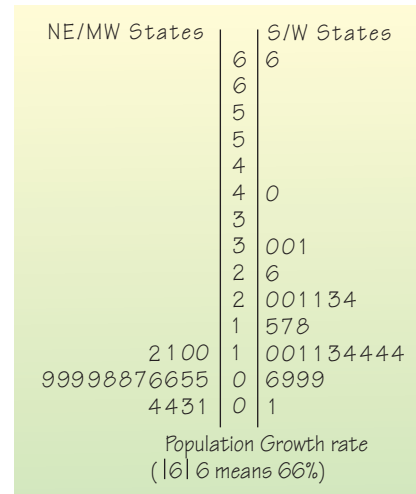
- a) What is the range of the sugar contents of these cereals.
- b) Describe the shape of the distribution.
- c) What aspect of breakfast cereals might account for this shape?
- d) Are all children’s cereals higher in sugar than adult cereals?
- e) Which group of cereals varies more in sugar content? Explain.

**T 10. Tendon transfers.** People with spinal cord injuries may lose function in some, but not all, of their muscles. The ability to push oneself up is particularly important for shifting position when seated and for transferring into and out of wheelchairs. Surgeons compared two operations to restore the ability to push up in children. The histogram shows scores rating pushing strength two years after surgery and boxplots compare results for the two surgical methods. (Mulcahey, Lutz, Kozen, Betz, “Prospective Evaluation of Biceps to Triceps and Deltoid to Triceps for Elbow Extension in Tetraplegia,” *Journal of Hand Surgery*, 28, 6, 2003)



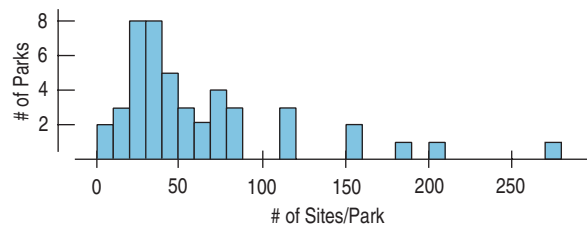
- a) Describe the shape of this distribution.
- b) What is the range of the strength scores?
- c) What fact about results of the two procedures is hidden in the histogram?
- d) Which method had the higher (better) median score?
- e) Was that method always best?
- f) Which method produced the most consistent results? Explain.

**T 11. Population growth.** Here is a “back-to-back” stem-and-leaf display that shows two data sets at once—one going to the left, one to the right. The display compares the percent change in population for two regions of the United States (based on census figures for 1990 and 2000). The fastest growing states were Nevada at 66% and Arizona at 40%. To show the distributions better, this display breaks each stem into two lines, putting leaves 0–4 on one stem and leaves 5–9 on the other.



- a) Use the data displayed in the stem-and-leaf display to construct comparative boxplots.
- b) Write a few sentences describing the difference in growth rates for the two regions of the United States.

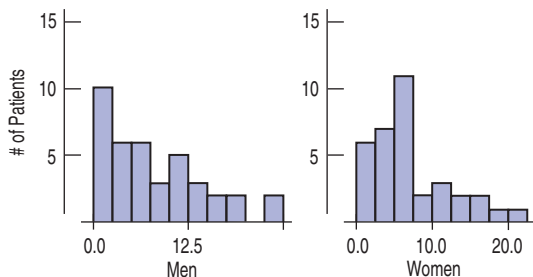
**12. Camp sites.** Shown below are the histogram and summary statistics for the number of camp sites at public parks in Vermont.



Count	46
Mean	62.8 sites
Median	43.5
StdDev	56.2
Min	0
Max	275
Q1	28
Q3	78

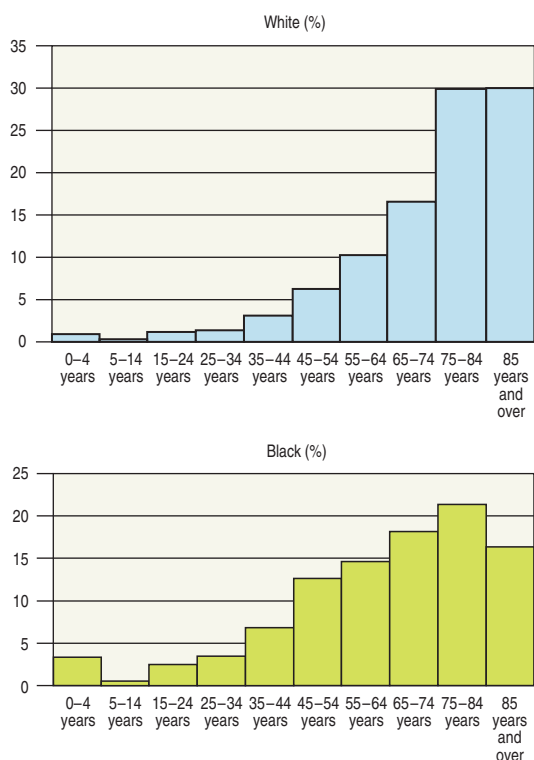
- a) Which statistics would you use to identify the center and spread of this distribution? Why?
- b) How many parks would you classify as outliers? Explain.
- c) Create a boxplot for these data.
- d) Write a few sentences describing the distribution.

13. **Hospital stays.** The U.S. National Center for Health Statistics compiles data on the length of stay by patients in short-term hospitals and publishes its findings in *Vital and Health Statistics*. Data from a sample of 39 male patients and 35 female patients on length of stay (in days) are displayed in the histograms below.



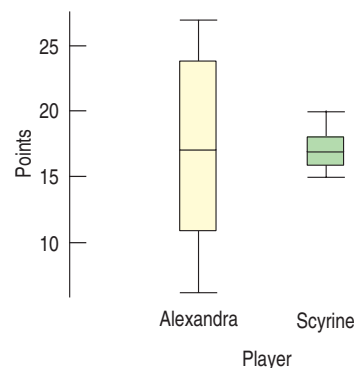
- a) What would you suggest be changed about these histograms to make them easier to compare?
- b) Describe these distributions by writing a few sentences comparing the duration of hospitalization for men and women.
- c) Can you suggest a reason for the peak in women's length of stay?

14. **Deaths 2003.** A National Vital Statistics Report ([www.cdc.gov/nchs/](http://www.cdc.gov/nchs/)) indicated that nearly 300,000 black Americans died in 2003, compared with just over 2 million white Americans. Here are histograms displaying the distributions of their ages at death:

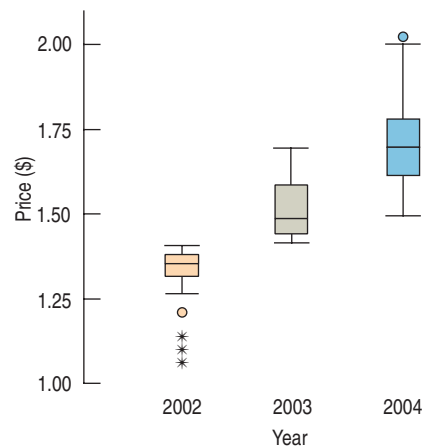


- a) Describe the overall shapes of these distributions.
- b) How do the distributions differ?
- c) Look carefully at the bar definitions. Where do these plots violate the rules for statistical graphs?

15. **Women's basketball.** Here are boxplots of the points scored during the first 10 games of the season for both Scyrine and Alexandra:

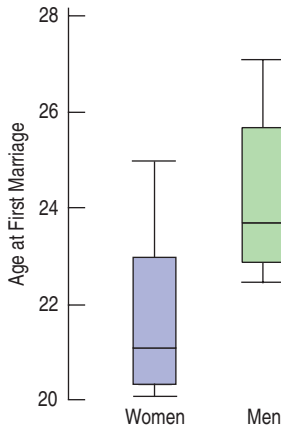


- a) Summarize the similarities and differences in their performance so far.
  - b) The coach can take only one player to the state championship. Which one should she take? Why?
16. **Gas prices.** Here are boxplots of weekly gas prices at a service station in the midwestern United States (prices in \$ per gallon):

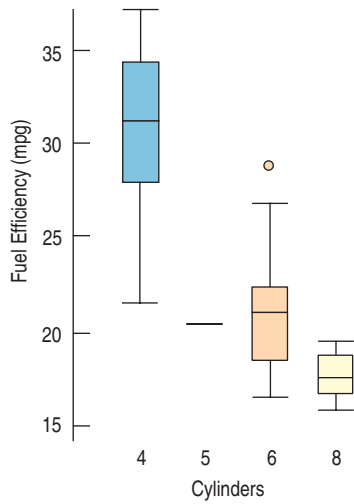


- a) Compare the distribution of prices over the three years.
- b) In which year were the prices least stable? Explain.

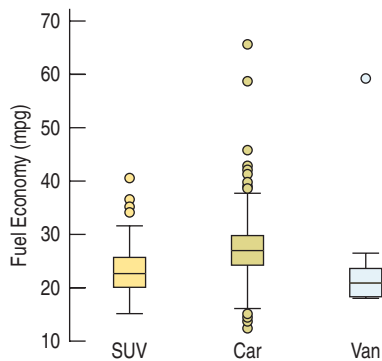
17. **Marriage age.** In 1975, did men and women marry at the same age? Here are boxplots of the age at first marriage for a sample of U.S. citizens then. Write a brief report discussing what these data show.



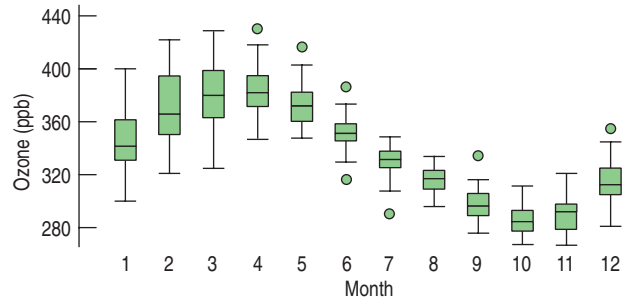
**T 18. Fuel economy.** Describe what these boxplots tell you about the relationship between the number of cylinders a car's engine has and the car's fuel economy (mpg):



**19. Fuel economy II.** The Environmental Protection Agency provides fuel economy and pollution information on over 2000 car models. Here is a boxplot of *Combined Fuel Economy* (using an average of driving conditions) in *miles per gallon* by vehicle *Type* (car, van, or SUV). Summarize what you see about the fuel economies of the three vehicle types.

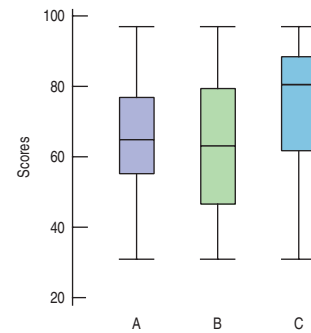
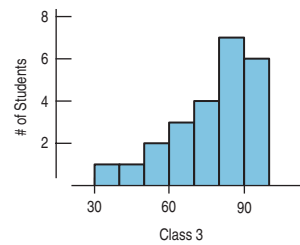
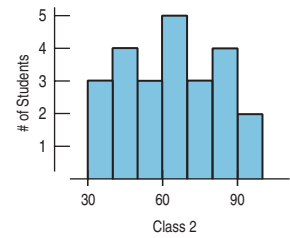
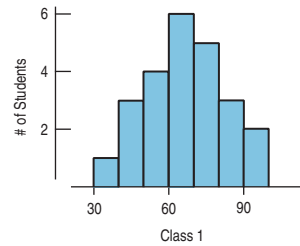


**T 20. Ozone.** Ozone levels (in parts per billion, ppb) were recorded at sites in New Jersey monthly between 1926 and 1971. Here are boxplots of the data for each month (over the 46 years), lined up in order (January = 1):



- In what month was the highest ozone level ever recorded?
- Which month has the largest IQR?
- Which month has the smallest range?
- Write a brief comparison of the ozone levels in January and June.
- Write a report on the annual patterns you see in the ozone levels.

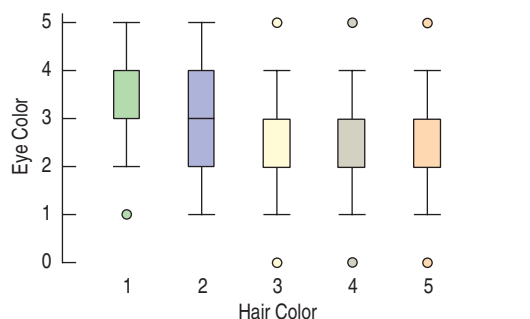
**21. Test scores.** Three Statistics classes all took the same test. Histograms and boxplots of the scores for each class are shown below. Match each class with the corresponding boxplot.



**22. Eye and hair color.** A survey of 1021 school-age children was conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey, the following codes were used:

Hair Color	Eye Color
1 = Blond	1 = Blue
2 = Brown	2 = Green
3 = Black	3 = Brown
4 = Red	4 = Grey
5 = Other	5 = Other

The Statistics students analyzing the data were asked to study the relationship between eye and hair color. They produced this plot:



Is their graph appropriate? If so, summarize the findings. If not, explain why not.

23. **Graduation?** A survey of major universities asked what percentage of incoming freshmen usually graduate “on time” in 4 years. Use the summary statistics given to answer the questions that follow.

	% on Time
Count	48
Mean	68.35
Median	69.90
StdDev	10.20
Min	43.20
Max	87.40
Range	44.20
25th %tile	59.15
75th %tile	74.75

- Would you describe this distribution as symmetric or skewed? Explain.
- Are there any outliers? Explain.
- Create a boxplot of these data.
- Write a few sentences about the graduation rates.

- T 24. **Vineyards.** Here are summary statistics for the sizes (in acres) of Finger Lakes vineyards:

Count	36
Mean	46.50 acres
StdDev	47.76
Median	33.50
IQR	36.50
Min	6
Q1	18.50
Q3	55
Max	250

- Would you describe this distribution as symmetric or skewed? Explain.
- Are there any outliers? Explain.
- Create a boxplot of these data.
- Write a few sentences about the sizes of the vineyards.

25. **Caffeine.** A student study of the effects of caffeine asked volunteers to take a memory test 2 hours after drinking soda. Some drank caffeine-free cola, some drank regular cola (with caffeine), and others drank a mixture of the two (getting a half-dose of caffeine). Here are the 5-number summaries for each group’s scores (number of items recalled correctly) on the memory test:

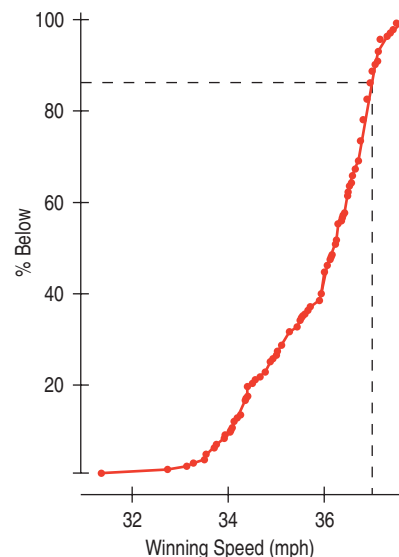
	<i>n</i>	Min	Q1	Median	Q3	Max
No caffeine	15	16	20	21	24	26
Low caffeine	15	16	18	21	24	27
High caffeine	15	12	17	19	22	24

- Describe the *W*’s for these data.
  - Name the variables and classify each as categorical or quantitative.
  - Create parallel boxplots to display these results as best you can with this information.
  - Write a few sentences comparing the performances of the three groups.
26. **SAT scores.** Here are the summary statistics for Verbal SAT scores for a high school graduating class:

	<i>n</i>	Mean	Median	SD	Min	Max	Q1	Q3
Male	80	590	600	97.2	310	800	515	650
Female	82	602	625	102.0	360	770	530	680

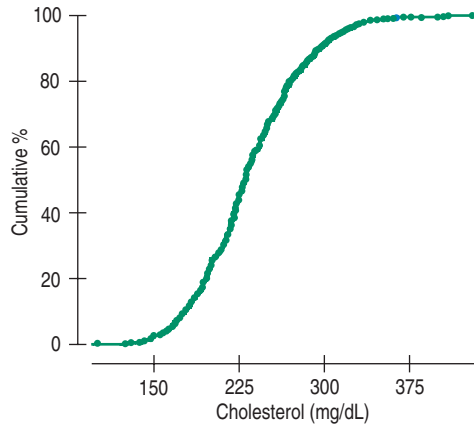
- Create parallel boxplots comparing the scores of boys and girls as best you can from the information given.
- Write a brief report on these results. Be sure to discuss the shape, center, and spread of the scores.

- T 27. **Derby speeds 2007.** How fast do horses run? Kentucky Derby winners top 30 miles per hour, as shown in this graph. The graph shows the percentage of Derby winners that have run *slower* than each given speed. Note that few have won running less than 33 miles per hour, but about 86% of the winning horses have run less than 37 miles per hour. (A cumulative frequency graph like this is called an “ogive.”)



- a) Estimate the median winning speed.
- b) Estimate the quartiles.
- c) Estimate the range and the IQR.
- d) Create a boxplot of these speeds.
- e) Write a few sentences about the speeds of the Kentucky Derby winners.

**T 28. Cholesterol.** The Framingham Heart Study recorded the cholesterol levels of more than 1400 men. Here is an ogive of the distribution of these cholesterol measures. (An ogive shows the percentage of cases at or below a certain value.) Construct a boxplot for these data, and write a few sentences describing the distribution.



**29. Reading scores.** A class of fourth graders takes a diagnostic reading test, and the scores are reported by reading grade level. The 5-number summaries for the 14 boys and 11 girls are shown:

**Boys:** 2.0 3.9 4.3 4.9 6.0  
**Girls:** 2.8 3.8 4.5 5.2 5.9

- a) Which group had the highest score?
- b) Which group had the greater range?
- c) Which group had the greater interquartile range?
- d) Which group's scores appear to be more skewed? Explain.
- e) Which group generally did better on the test? Explain.
- f) If the mean reading level for boys was 4.2 and for girls was 4.6, what is the overall mean for the class?

**T 30. Rainmakers?** In an experiment to determine whether seeding clouds with silver iodide increases rainfall, 52 clouds were randomly assigned to be seeded or not. The amount of rain they generated was then measured (in acre-feet). Here are the summary statistics:

	<i>n</i>	Mean	Median	SD	IQR	Q1	Q3
Unseeded	26	164.59	44.20	278.43	138.60	24.40	163
Seeded	26	441.98	221.60	650.79	337.60	92.40	430

- a) Which of the summary statistics are most appropriate for describing these distributions. Why?
- b) Do you see any evidence that seeding clouds may be effective? Explain.

**T 31. Industrial experiment.** Engineers at a computer production plant tested two methods for accuracy in drilling holes into a PC board. They tested how fast they could set the drilling machine by running 10 boards at each of two different speeds. To assess the results, they measured the distance (in inches) from the center of a target on the board to the center of the hole. The data and summary statistics are shown in the table:

	Distance (in.)	Speed		Distance (in.)	Speed
	0.000101	Fast		0.000098	Slow
	0.000102	Fast		0.000096	Slow
	0.000100	Fast		0.000097	Slow
	0.000102	Fast		0.000095	Slow
	0.000101	Fast		0.000094	Slow
	0.000103	Fast		0.000098	Slow
	0.000104	Fast		0.000096	Slow
	0.000102	Fast		0.975600	Slow
	0.000102	Fast		0.000097	Slow
	0.000100	Fast		0.000096	Slow
Mean	0.000102		Mean	0.097647	
StdDev	0.000001		StdDev	0.308481	

Write a report summarizing the findings of the experiment. Include appropriate visual and verbal displays of the distributions, and make a recommendation to the engineers if they are most interested in the accuracy of the method.

**T 32. Cholesterol.** A study examining the health risks of smoking measured the cholesterol levels of people who had smoked for at least 25 years and people of similar ages who had smoked for no more than 5 years and then stopped. Create appropriate graphical displays for both groups, and write a brief report comparing their cholesterol levels. Here are the data:

Smokers				Ex-Smokers		
225	211	209	284	250	134	300
258	216	196	288	249	213	310
250	200	209	280	175	174	328
225	256	243	200	160	188	321
213	246	225	237	213	257	292
232	267	232	216	200	271	227
216	243	200	155	238	163	263
216	271	230	309	192	242	249
183	280	217	305	242	267	243
287	217	246	351	217	267	218
200	280	209		217	183	228

**T 33. MPG.** A consumer organization compared gas mileage figures for several models of cars made in the United States with autos manufactured in other countries. The data are shown in the table:



Gas Mileage (mpg)	Country	Gas Mileage (mpg)	Country
16.9	U.S.	26.8	U.S.
15.5	U.S.	33.5	U.S.
19.2	U.S.	34.2	U.S.
18.5	U.S.	16.2	Other
30.0	U.S.	20.3	Other
30.9	U.S.	31.5	Other
20.6	U.S.	30.5	Other
20.8	U.S.	21.5	Other
18.6	U.S.	31.9	Other
18.1	U.S.	37.3	Other
17.0	U.S.	27.5	Other
17.6	U.S.	27.2	Other
16.5	U.S.	34.1	Other
18.2	U.S.	35.1	Other
26.5	U.S.	29.5	Other
21.9	U.S.	31.8	Other
27.4	U.S.	22.0	Other
28.4	U.S.	17.0	Other
28.8	U.S.	21.6	Other

- a) Create graphical displays for these two groups.
- b) Write a few sentences comparing the distributions.

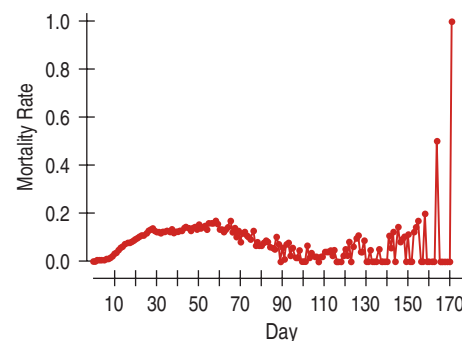
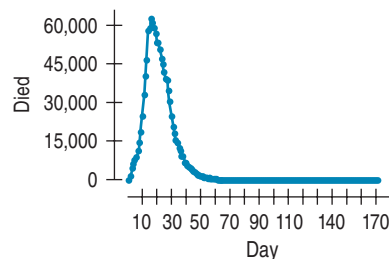
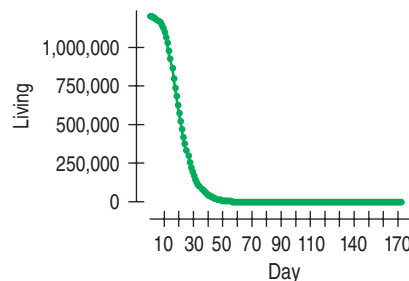
**T 34. Baseball.** American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The League believes that replacing the pitcher, typically a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Following are the average number of runs scored in American League and National League stadiums for the first half of the 2001 season:

Average Runs	League	Average Runs	League
11.1	American	14.0	National
10.8	American	11.6	National
10.8	American	10.4	National
10.3	American	10.9	National
10.3	American	10.2	National
10.1	American	9.5	National
10.0	American	9.5	National
9.5	American	9.5	National
9.4	American	9.5	National
9.3	American	9.1	National
9.2	American	8.8	National
9.2	American	8.4	National
9.0	American	8.3	National
8.3	American	8.2	National
		8.1	National
		7.9	National

- a) Create an appropriate graphical display of these data.
- b) Write a few sentences comparing the average number of runs scored per game in the two leagues. (Remember: shape, center, spread, unusual features!)

c) Coors Field in Denver stands a mile above sea level, an altitude far greater than that of any other major league ball park. Some believe that the thinner air makes it harder for pitchers to throw curveballs and easier for batters to hit the ball a long way. Do you see any evidence that the 14 runs scored per game there is unusually high? Explain.

**T 35. Fruit Flies.** Researchers tracked a population of 1,203,646 fruit flies, counting how many died each day for 171 days. Here are three timeplots offering different views of these data. One shows the number of flies alive on each day, one the number who died that day, and the third the mortality rate—the fraction of the number alive who died. On the last day studied, the last 2 flies died, for a mortality rate of 1.0.



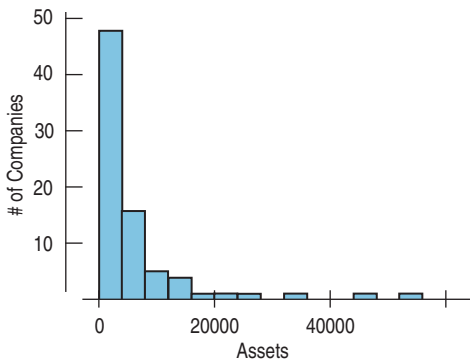
- a) On approximately what day did the most flies die?
- b) On what day during the first 100 days did the largest proportion of flies die?
- c) When did the number of fruit flies alive stop changing very much from day to day?

**T 36. Drunk driving 2005.** Accidents involving drunk drivers account for about 40% of all deaths on the nation's highways. The table tracks the number of alcohol-related fatalities for 24 years. ([www.madd.org](http://www.madd.org))

Year	Deaths (thousands)	Year	Deaths (thousands)
1982	26.2	1994	17.3
1983	24.6	1995	17.7
1984	24.8	1996	17.7
1985	23.2	1997	16.7
1986	25.0	1998	16.7
1987	24.1	1999	16.6
1988	23.8	2000	17.4
1989	22.4	2001	17.4
1990	22.6	2002	17.5
1991	20.2	2003	17.1
1992	18.3	2004	16.9
1993	17.9	2005	16.9

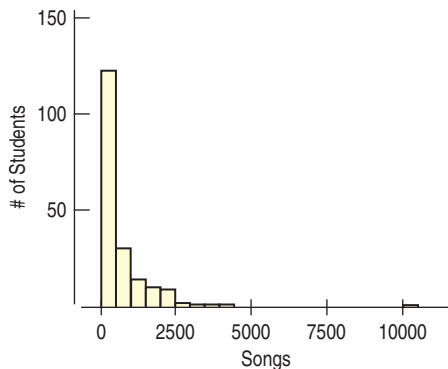
- Create a stem-and-leaf display or a histogram of these data.
- Create a timeplot.
- Using features apparent in the stem-and-leaf display (or histogram) and the timeplot, write a few sentences about deaths caused by drunk driving.

**T 37. Assets.** Here is a histogram of the assets (in millions of dollars) of 79 companies chosen from the *Forbes* list of the nation's top corporations:



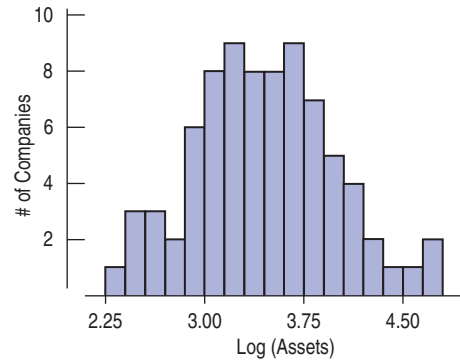
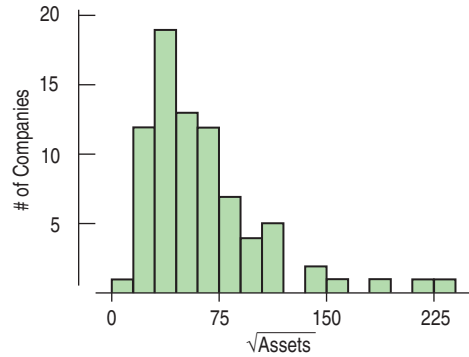
- What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
- What would you suggest doing with these data if we want to understand them better?

**38. Music library.** Students were asked how many songs they had in their digital music libraries. Here's a display of the responses:



- What aspect of this distribution makes it difficult to summarize, or to discuss, center and spread?
- What would you suggest doing with these data if we want to understand them better?

**T 39. Assets again.** Here are the same data you saw in Exercise 37 after re-expressions as the square root of assets and the logarithm of assets:



- Which re-expression do you prefer? Why?
- In the square root re-expression, what does the value 50 actually indicate about the company's assets?
- In the logarithm re-expression, what does the value 3 actually indicate about the company's assets?

**T 40. Rainmakers.** The table lists the amount of rainfall (in acre-feet) from the 26 clouds seeded with silver iodide discussed in Exercise 30:

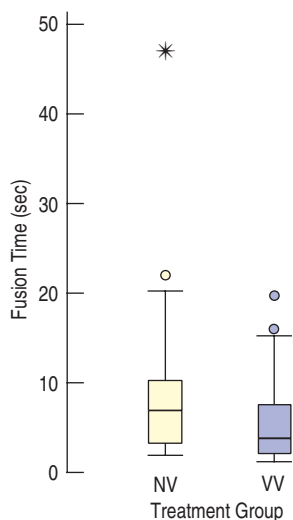
2745	703	302	242	119	40	7
1697	489	274	200	118	32	4
1656	430	274	198	115	31	
978	334	255	129	92	17	

- Why is acre-feet a good way to measure the amount of precipitation produced by cloud seeding?
- Plot these data, and describe the distribution.
- Create a re-expression of these data that produces a more advantageous distribution.
- Explain what your re-expressed scale means.

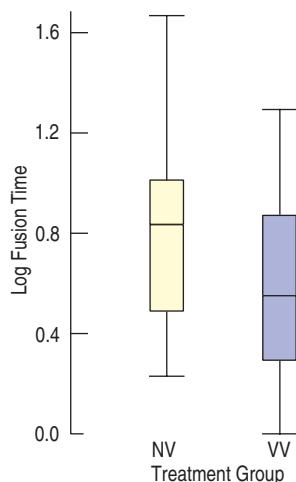
**T 41. Stereograms.** Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the embedded image affected the

time required for subjects to fuse the images. One group of subjects (group NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

- What two variables are discussed in this description?
- For each variable, is it quantitative or categorical? If quantitative, what are the units?
- The boxplots compare the fusion times for the two treatment groups. Write a few sentences comparing these distributions. What does the experiment show?



- T 42. Stereograms, revisited.** Because of the skewness of the distributions of fusion times described in Exercise 41, we might consider a re-expression. Here are the boxplots of the  $\log$  of fusion times. Is it better to analyze the original fusion times or the log fusion times? Explain.



### JUST CHECKING Answers

- The % late arrivals have a unimodal, symmetric distribution centered at about 20%. In most months between 16% and 23% of the flights arrived late.
- The boxplot of % late arrivals makes it easier to see that the median is just below 20%, with quartiles at about 17% and 22%. It nominates two months as high outliers.
- The boxplots by month show a strong seasonal pattern. Flights are more likely to be late in the winter and summer and less likely to be late in the spring and fall. One likely reason for the pattern is snowstorms in the winter and thunderstorms in the summer.

# The Standard Deviation as a Ruler and the Normal Model



The women's heptathlon in the Olympics consists of seven track and field events: the 200-m and 800-m runs, 100-m high hurdles, shot put, javelin, high jump, and long jump. To determine who should get the gold medal, somehow the performances in all seven events have to be combined into one score. How can performances in such different events be compared? They don't even have the same units; the races are recorded in minutes and seconds and the throwing and jumping events in meters. In the 2004 Olympics, Austra Skujytė of Lithuania put the shot 16.4 meters, about 3 meters farther than the average of all contestants. Carolina Klüft won the long jump with a 6.78-m jump, about a meter better than the average. Which performance deserves more points? Even though both events are measured in meters, it's not clear how to compare them. The solution to the problem of how to compare scores turns out to be a useful method for comparing all sorts of values whether they have the same units or not.

## The Standard Deviation as a Ruler

### Grading on a Curve

If you score 79% on an exam, what grade should you get? One teaching philosophy looks only at the raw percentage, 79, and bases the grade on that alone. Another looks at your *relative* performance and bases the grade on how you did compared with the rest of the class. Teachers and students still debate which method is better.

The trick in comparing very different-looking values is to use standard deviations. The standard deviation tells us how the whole collection of values varies, so it's a natural ruler for comparing an individual value to the group. Over and over during this course, we will ask questions such as "How far is this value from the mean?" or "How different are these two statistics?" The answer in every case will be to measure the distance or difference in standard deviations.

The concept of the standard deviation as a ruler is not special to this course. You'll find statistical distances measured in standard deviations throughout Statistics, up to the most advanced levels.<sup>1</sup> This approach is one of the basic tools of statistical thinking.

<sup>1</sup> Other measures of spread could be used as well, but the standard deviation is the most common measure, and it is almost always used as the ruler.

In order to compare the two events, let's start with a picture. This time we'll use stem-and-leaf displays so we can see the individual distances.

Long Jump		Shot Put	
Stem	Leaf	Stem	Leaf
67	8	16	4
66		15	
65	1	15	
64	2	14	56778
63	0566	14	24
62	11235	13	5789
61	0569	13	012234
60	2223	12	55
59	0278	12	0144
58	4	11	59
57	0	11	23

FIGURE 6.1

Stem-and-leaf displays for both the long jump and the shot put in the 2004 Olympic Heptathlon. Carolina Klüft (green scores) won the long jump, and Austra Skujytė (red scores) won the shot put. Which heptathlete did better for both events combined?

The two winning performances on the top of each stem-and-leaf display appear to be about the same distance from the center of the pack. But look again carefully. What do we mean by the *same distance*? The two displays have different scales. Each line in the stem-and-leaf for the shot put represents half a meter, but for the long jump each line is only a tenth of a meter. It's only because our eyes naturally adjust the scales and use the standard deviation as the ruler that we see each as being about the same distance from the center of the data. How can we make this hunch more precise? Let's see how many standard deviations each performance is from the mean.

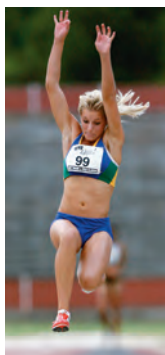
Klüft's 6.78-m long jump is 0.62 meters longer than the mean jump of 6.16 m. How many *standard deviations* better than the mean is that? The standard deviation for this event was 0.23 m, so her jump was  $(6.78 - 6.16)/0.23 = 0.62/0.23 = 2.70$  *standard deviations* better than the mean. Skujytė's winning shot put was  $16.40 - 13.29 = 3.11$  meters longer than the mean shot put distance, and that's  $3.11/1.24 = 2.51$  standard deviations better than the mean. That's a great performance but not quite as impressive as Klüft's long jump, which was farther above the mean, as measured in *standard deviations*.

	Event	
	Long Jump	Shot Put
Mean (all contestants)	6.16 m	13.29 m
SD	0.23 m	1.24 m
$n$	26	28
Klüft	6.78 m	14.77 m
Skujytė	6.30 m	16.40 m

## Standardizing with z-Scores

### NOTATION ALERT:

There goes another letter. We always use the letter  $z$  to denote values that have been standardized with the mean and standard deviation.



To compare these athletes' performances, we determined how many standard deviations from the event's mean each was.

Expressing the distance in standard deviations *standardizes* the performances. To standardize a value, we simply subtract the mean performance in that event and then divide this difference by the standard deviation. We can write the calculation as

$$z = \frac{y - \bar{y}}{s}$$

These values are called **standardized values**, and are commonly denoted with the letter  $z$ . Usually, we just call them **z-scores**.

Standardized values have *no units*. z-scores measure the distance of each data value from the mean in standard deviations. A z-score of 2 tells us that a data value is 2 standard deviations above the mean. It doesn't matter whether the original variable was measured in inches, dollars, or seconds. Data values below the mean have negative z-scores, so a z-score of  $-1.6$  means that the data value was 1.6 standard deviations below the mean. Of course, regardless of the direction, the farther a data value is from the mean, the more unusual it is, so a z-score of  $-1.3$

is more extraordinary than a z-score of 1.2. Looking at the z-scores, we can see that even though both were winning scores, Klüft's long jump with a z-score of 2.70 is slightly more impressive than Skujyté's shot put with a z-score of 2.51.

**FOR EXAMPLE**

**Standardizing skiing times**

The men's combined skiing event in the winter Olympics consists of two races: a downhill and a slalom. Times for the two events are added together, and the skier with the lowest total time wins. In the 2006 Winter Olympics, the mean slalom time was 94.2714 seconds with a standard deviation of 5.2844 seconds. The mean downhill time was 101.807 seconds with a standard deviation of 1.8356 seconds. Ted Ligety of the United States, who won the gold medal with a combined time of 189.35 seconds, skied the slalom in 87.93 seconds and the downhill in 101.42 seconds.

**Question:** On which race did he do better compared with the competition?

For the slalom, Ligety's z-score is found by subtracting the mean time from his time and then dividing by the standard deviation:

$$z_{\text{Slalom}} = \frac{87.93 - 94.2714}{5.2844} = -1.2$$

Similarly, his z-score for the downhill is:

$$z_{\text{Downhill}} = \frac{101.42 - 101.807}{1.8356} = -0.21$$

The z-scores show that Ligety's time in the slalom is farther below the mean than his time in the downhill. His performance in the slalom was more remarkable.

By using the standard deviation as a ruler to measure statistical distance from the mean, we can compare values that are measured on different variables, with different scales, with different units, or for different individuals. To determine the winner of the heptathlon, the judges must combine performances on seven very different events. Because they want the score to be absolute, and *not* dependent on the particular athletes in each Olympics, they use predetermined tables, but they could combine scores by standardizing each, and then adding the z-scores together to reach a total score. The only trick is that they'd have to switch the sign of the z-score for running events, because unlike throwing and jumping, it's better to have a running time below the mean (with a negative z-score).

To combine the scores Skujyté and Klüft earned in the long jump and the shot put, we standardize both events as shown in the table. That gives Klüft her 2.70 z-score in the long jump and a 1.19 in the shot put, for a total of 3.89. Skujyté's shot put gave her a 2.51, but her long jump was only 0.61 SDs above the mean, so her total is 3.12.

		Event	
		Long Jump	Shot Put
	Mean	6.16 m	13.29 m
	SD	0.23 m	1.24 m
Klüft	Performance	6.78 m	14.77 m
	z-score	$\frac{6.78 - 6.16}{0.23} = 2.70$	$\frac{14.77 - 13.29}{1.24} = 1.19$
	Total z-score	2.70 + 1.19 = <b>3.89</b>	
Skujyté	Performance	6.30 m	16.40 m
	z-score	$\frac{6.30 - 6.16}{0.23} = 0.61$	$\frac{16.40 - 13.29}{1.24} = 2.51$
	Total z-score	0.61 + 2.51 = <b>3.12</b>	

Is this the result we wanted? Yes. Each won one event, but Klüft's shot put was second best, while Skujyté's long jump was seventh. The z-scores measure how far each result is from the event mean in standard deviation units. And because they are both in standard deviation units, we can combine them. Not coincidentally, Klüft went on to win the gold medal for the entire seven-event heptathlon, while Skujyté got the silver.

## FOR EXAMPLE

## Combining z-scores

In the 2006 winter Olympics men's combined event, Ivica Kostelić of Croatia skied the slalom in 89.44 seconds and the downhill in 100.44 seconds. He thus beat Ted Ligety in the downhill, but not in the slalom. Maybe he should have won the gold medal.

**Question:** Considered in terms of standardized scores, which skier did better?

Kostelić's z-scores are:

$$z_{\text{Slalom}} = \frac{89.44 - 94.2714}{5.2844} = -0.91 \quad \text{and} \quad z_{\text{Downhill}} = \frac{100.44 - 101.807}{1.8356} = -0.74$$

The sum of his z-scores is approximately  $-1.65$ . Ligety's z-score sum is only about  $-1.41$ . Because the standard deviation of the downhill times is so much smaller, Kostelić's better performance there means that he would have won the event if standardized scores were used.

When we standardize data to get a z-score, we do two things. First, we shift the data by subtracting the mean. Then, we rescale the values by dividing by their standard deviation. We often shift and rescale data. What happens to a grade distribution if *everyone* gets a five-point bonus? Everyone's grade goes up, but does the shape change? (*Hint:* Has anyone's distance from the mean changed?) If we switch from feet to meters, what happens to the distribution of heights of students in your class? Even though your intuition probably tells you the answers to these questions, we need to look at exactly how shifting and rescaling work.



## JUST CHECKING

1. Your Statistics teacher has announced that the lower of your two tests will be dropped. You got a 90 on test 1 and an 80 on test 2. You're all set to drop the 80 until she announces that she grades "on a curve." She standardized the scores in order to decide which is the lower one. If the mean on the first test was 88 with a standard deviation of 4 and the mean on the second was 75 with a standard deviation of 5,
  - a) Which one will be dropped?
  - b) Does this seem "fair"?

## Shifting Data

Since the 1960s, the Centers for Disease Control's National Center for Health Statistics has been collecting health and nutritional information on people of all ages and backgrounds. A recent survey, the National Health and Nutrition Examination Survey (NHANES) 2001–2002,<sup>2</sup> measured a wide variety of variables, including body measurements, cardiovascular fitness, blood chemistry, and demographic information on more than 11,000 individuals.

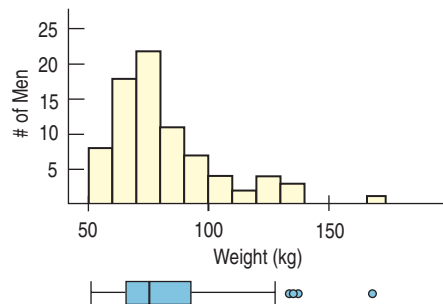
<sup>2</sup> [www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm)

<b>WHO</b>	80 male participants of the NHANES survey between the ages of 19 and 24 who measured between 68 and 70 inches tall
<b>WHAT</b>	Their weights
<b>UNIT</b>	Kilograms
<b>WHEN</b>	2001–2002
<b>WHERE</b>	United States
<b>WHY</b>	To study nutrition, and health issues and trends
<b>HOW</b>	National survey

**AS** **Activity: Changing the Baseline.** What happens when we shift data? Do measures of center and spread change?

Doctors' height and weight charts sometimes give ideal weights for various heights that include 2-inch heels. If the mean height of adult women is 66 inches including 2-inch heels, what is the mean height of women without shoes? Each woman is shorter by 2 inches when barefoot, so the mean is decreased by 2 inches, to 64 inches.

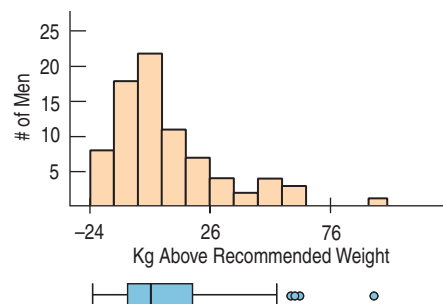
Included in this group were 80 men between 19 and 24 years old of average height (between 5'8" and 5'10" tall). Here are a histogram and boxplot of their weights:



**FIGURE 6.2**

*Histogram and boxplot for the men's weights. The shape is skewed to the right with several high outliers.*

Their mean weight is 82.36 kg. For this age and height group, the National Institutes of Health recommends a maximum healthy weight of 74 kg, but we can see that some of the men are heavier than the recommended weight. To compare their weights to the recommended maximum, we could subtract 74 kg from each of their weights. What would that do to the center, shape, and spread of the histogram? Here's the picture:



**FIGURE 6.3**

*Subtracting 74 kilograms shifts the entire histogram down but leaves the spread and the shape exactly the same.*

On average, they weigh 82.36 kg, so on average they're 8.36 kg overweight. And, after subtracting 74 from each weight, the mean of the new distribution is  $82.36 - 74 = 8.36$  kg. In fact, when we **shift** the data by adding (or subtracting) a constant to each value, all measures of position (center, percentiles, min, max) will increase (or decrease) by the same constant.

What about the spread? What does adding or subtracting a constant value do to the spread of the distribution? Look at the two histograms again. Adding or subtracting a constant changes each data value equally, so the entire distribution just shifts. Its shape doesn't change and neither does the spread. None of the measures of spread we've discussed—not the range, not the IQR, not the standard deviation—changes.

*Adding (or subtracting) a constant to every data value adds (or subtracts) the same constant to measures of position, but leaves measures of spread unchanged.*

## Rescaling Data

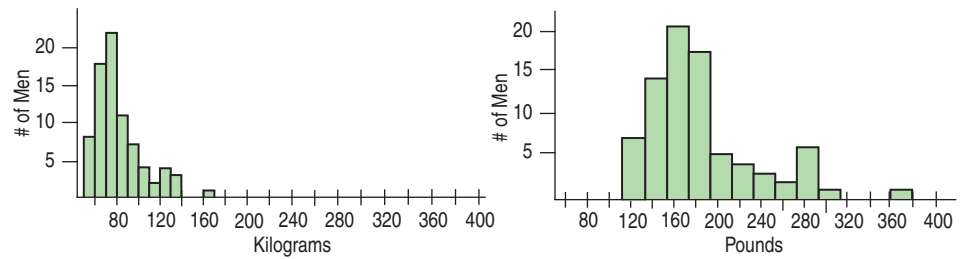
Not everyone thinks naturally in metric units. Suppose we want to look at the weights in pounds instead. We'd have to **rescale** the data. Because there are about 2.2 pounds in every kilogram, we'd convert the weights by multiplying each value by 2.2. Multiplying or dividing each value by a constant changes the measurement



units. Here are histograms of the two weight distributions, plotted on the same scale, so you can see the effect of multiplying:

**FIGURE 6.4**

*Men's weights in both kilograms and pounds. How do the distributions and numerical summaries change?*



**A S** **Simulation: Changing the Units.** Change the center and spread values for a distribution and watch the summaries change (or not, as the case may be).

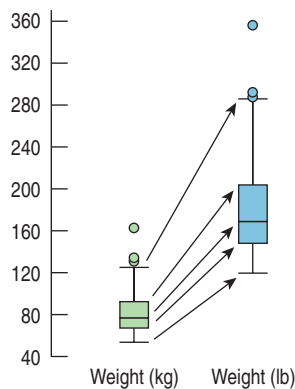
What happens to the shape of the distribution? Although the histograms don't look exactly alike, we see that the shape really hasn't changed: Both are unimodal and skewed to the right.

What happens to the mean? Not too surprisingly, it gets multiplied by 2.2 as well. The men weigh 82.36 kg on average, which is 181.19 pounds. As the boxplots and 5-number summaries show, all measures of position act the same way. They all get multiplied by this same constant.

What happens to the spread? Take a look at the boxplots. The spread in pounds (on the right) is larger. How much larger? If you guessed 2.2 times, you've figured out how measures of spread get rescaled.

**FIGURE 6.5**

*The boxplots (drawn on the same scale) show the weights measured in kilograms (on the left) and pounds (on the right). Because 1 kg is 2.2 lb, all the points in the right box are 2.2 times larger than the corresponding points in the left box. So each measure of position and spread is 2.2 times as large when measured in pounds rather than kilograms.*



	Weight (kg)	Weight (lb)
Min	54.3	119.46
Q1	67.3	148.06
Median	76.85	169.07
Q3	92.3	203.06
Max	161.5	355.30
IQR	25	55
SD	22.27	48.99

*When we multiply (or divide) all the data values by any constant, all measures of position (such as the mean, median, and percentiles) and measures of spread (such as the range, the IQR, and the standard deviation) are multiplied (or divided) by that same constant.*

## FOR EXAMPLE

### Rescaling the slalom

**Recap:** The times in the men's combined event at the winter Olympics are reported in minutes and seconds. Previously, we converted these to seconds and found the mean and standard deviation of the slalom times to be 94.2714 seconds and 5.2844 seconds, respectively.

**Question:** Suppose instead that we had reported the times in minutes—that is, that each individual time was divided by 60. What would the resulting mean and standard deviation be?

Dividing all the times by 60 would divide both the mean and the standard deviation by 60:

$$\text{Mean} = 94.2714/60 = 1.5712 \text{ minutes}; \quad \text{SD} = 5.2844/60 = 0.0881 \text{ minutes}.$$



## JUST CHECKING

2. In 1995 the Educational Testing Service (ETS) adjusted the scores of SAT tests. Before ETS recentered the SAT Verbal test, the mean of all test scores was 450.
  - a) How would adding 50 points to each score affect the mean?
  - b) The standard deviation was 100 points. What would the standard deviation be after adding 50 points?
  - c) Suppose we drew boxplots of test takers' scores a year before and a year after the recentering. How would the boxplots of the two years differ?
3. A company manufactures wheels for in-line skates. The diameters of the wheels have a mean of 3 inches and a standard deviation of 0.1 inches. Because so many of their customers use the metric system, the company decided to report their production statistics in millimeters (1 inch = 25.4 mm). They report that the standard deviation is now 2.54 mm. A corporate executive is worried about this increase in variation. Should he be concerned? Explain.

## Back to z-scores

**AS**

### Activity: Standardizing.

What if we both shift and rescale?  
The result is so nice that we give  
it a name.

z-scores have mean 0 and  
standard deviation 1.

Standardizing data into z-scores is just shifting them by the mean and rescaling them by the standard deviation. Now we can see how standardizing affects the distribution. When we subtract the mean of the data from every data value, we shift the mean to zero. As we have seen, such a shift doesn't change the standard deviation.

When we *divide* each of these shifted values by  $s$ , however, the standard deviation should be divided by  $s$  as well. Since the standard deviation was  $s$  to start with, the new standard deviation becomes 1.

How, then, does standardizing affect the distribution of a variable? Let's consider the three aspects of a distribution: the shape, center, and spread.

- ▶ Standardizing into z-scores does not change the **shape** of the distribution of a variable.
- ▶ Standardizing into z-scores changes the **center** by making the mean 0.
- ▶ Standardizing into z-scores changes the **spread** by making the standard deviation 1.




### STEP-BY-STEP EXAMPLE

### Working with Standardized Variables

Many colleges and universities require applicants to submit scores on standardized tests such as the SAT Writing, Math, and Critical Reading (Verbal) tests. The college your little sister wants to apply to says that while there is no minimum score required, the middle 50% of their students have combined SAT scores between 1530 and 1850. You'd feel confident if you knew her score was in their top 25%, but unfortunately she took the ACT test, an alternative standardized test.

**Question:** How high does her ACT need to be to make it into the top quarter of equivalent SAT scores?

To answer that question you'll have to standardize all the scores, so you'll need to know the mean and standard deviations of scores for some group on both tests. The college doesn't report the mean or standard deviation for their applicants on either test, so we'll use the group of all test takers nationally. For college-bound seniors, the average combined SAT score is about 1500 and the standard deviation is about 250 points. For the same group, the ACT average is 20.8 with a standard deviation of 4.8.

 <p><b>Plan</b> State what you want to find out.</p> <p><b>Variables</b> Identify the variables and report the W's (if known).</p> <p>Check the appropriate conditions.</p>	<p>I want to know what ACT score corresponds to the upper-quartile SAT score. I know the mean and standard deviation for both the SAT and ACT scores based on all test takers, but I have no individual data values.</p> <p>✓ <b>Quantitative Data Condition:</b> Scores for both tests are quantitative but have no meaningful units other than points.</p>
 <p><b>Mechanics</b> Standardize the variables.</p> <p>The <math>y</math>-value we seek is <math>z</math> standard deviations above the mean.</p>	<p>The middle 50% of SAT scores at this college fall between 1530 and 1850 points. To be in the top quarter, my sister would have to have a score of at least 1850. That's a <math>z</math>-score of</p> $z = \frac{(1850 - 1500)}{250} = 1.40$ <p>So an SAT score of 1850 is 1.40 standard deviations above the mean of all test takers.</p> <p>For the ACT, 1.40 standard deviations above the mean is <math>20.8 + 1.40(4.8) = 27.52</math>.</p>
 <p><b>Conclusion</b> Interpret your results in context.</p>	<p>To be in the top quarter of applicants in terms of combined SAT score, she'd need to have an ACT score of at least 27.52.</p>

## When Is a z-score BIG?

A  $z$ -score gives us an indication of how unusual a value is because it tells us how far it is from the mean. If the data value sits right at the mean, it's not very far at all and its  $z$ -score is 0. A  $z$ -score of 1 tells us that the data value is 1 standard deviation above the mean, while a  $z$ -score of  $-1$  tells us that the value is 1 standard deviation below the mean. How far from 0 does a  $z$ -score have to be to be interesting or unusual? There is no universal standard, but the larger the score is (negative or positive), the more unusual it is. We know that 50% of the data lie between the quartiles. For symmetric data, the standard deviation is usually a bit smaller than the IQR, and it's not uncommon for at least half of the data to have  $z$ -scores between  $-1$  and 1. But no matter what the shape of the distribution, a  $z$ -score of 3 (plus or minus) or more is rare, and a  $z$ -score of 6 or 7 shouts out for attention.

To say more about how big we expect a  $z$ -score to be, we need to *model* the data's distribution. A model will let us say much more precisely how often we'd be likely to see  $z$ -scores of different sizes. Of course, like all models of the real world, the model will be wrong—wrong in the sense that it can't match

### Is Normal Normal?

Don't be misled. The name "Normal" doesn't mean that these are the *usual* shapes for histograms. The name follows a tradition of positive thinking in Mathematics and Statistics in which functions, equations, and relationships that are easy to work with or have other nice properties are called "normal", "common", "regular", "natural", or similar terms. It's as if by calling them ordinary, we could make them actually occur more often and simplify our lives.

“All models are wrong—but some are useful.”

—George Box, famous statistician

### NOTATION ALERT:

$N(\mu, \sigma)$  always denotes a Normal model. The  $\mu$ , pronounced “mew,” is the Greek letter for “m” and always represents the mean in a model. The  $\sigma$ , sigma, is the lowercase Greek letter for “s” and always represents the standard deviation in a model.

### Is the Standard Normal a standard?

Yes. We call it the “Standard Normal” because it models standardized values. It is also a “standard” because this is the particular Normal model that we almost always use.

**A S** **Activity: Working with Normal Models.** Learn more about the Normal model and see what data drawn at random from a Normal model might look like.

reality exactly. But it can still be useful. Like a physical model, it’s something we can look at and manipulate in order to learn more about the real world.

Models help our understanding in many ways. Just as a model of an airplane in a wind tunnel can give insights even though it doesn’t show every rivet,<sup>3</sup> models of data give us summaries that we can learn from and use, even though they don’t fit each data value exactly. It’s important to remember that they’re only *models* of reality and not reality itself. But without models, what we can learn about the world at large is limited to only what we can say about the data we have at hand.

There is no universal standard for z-scores, but there is a model that shows up over and over in Statistics. You may have heard of “bell-shaped curves.” Statisticians call them Normal models. **Normal models** are appropriate for distributions whose shapes are unimodal and roughly symmetric. For these distributions, they provide a measure of how extreme a z-score is. Fortunately, there is a Normal model for every possible combination of mean and standard deviation. We write  $N(\mu, \sigma)$  to represent a Normal model with a mean of  $\mu$  and a standard deviation of  $\sigma$ . Why the Greek? Well, *this* mean and standard deviation are not numerical summaries of data. They are part of the model. They don’t come from the data. Rather, they are numbers that we choose to help specify the model. Such numbers are called **parameters** of the model.

We don’t want to confuse the parameters with summaries of the data such as  $\bar{y}$  and  $s$ , so we use special symbols. In Statistics, we almost always use Greek letters for parameters. By contrast, summaries of data are called **statistics** and are usually written with Latin letters.

If we model data with a Normal model and standardize them using the corresponding  $\mu$  and  $\sigma$ , we still call the standardized value a **z-score**, and we write

$$z = \frac{y - \mu}{\sigma}.$$

Usually it’s easier to standardize data first (using its mean and standard deviation). Then we need only the model  $N(0,1)$ . The Normal model with mean 0 and standard deviation 1 is called the **standard Normal model** (or the **standard Normal distribution**).

But be careful. You shouldn’t use a Normal model for just any data set. Remember that standardizing won’t change the shape of the distribution. If the distribution is not unimodal and symmetric to begin with, standardizing won’t make it Normal.

When we use the Normal model, we assume that the distribution of the data is, well, Normal. Practically speaking, there’s no way to check whether this **Normality Assumption** is true. In fact, it almost certainly is not true. Real data don’t behave like mathematical models. Models are idealized; real data are real. The good news, however, is that to use a Normal model, it’s sufficient to check the following condition:

**Nearly Normal Condition.** The shape of the data’s distribution is unimodal and symmetric. Check this by making a histogram (or a Normal probability plot, which we’ll explain later).

Don’t model data with a Normal model without checking whether the condition is satisfied.

All models make **assumptions**. Whenever we model—and we’ll do that often—we’ll be careful to point out the assumptions that we’re making. And, what’s even more important, we’ll check the associated **conditions** in the data to make sure that those assumptions are reasonable.

<sup>3</sup> In fact, the model is useful *because* it doesn’t have every rivet. It is because models offer a simpler view of reality that they are so useful as we try to understand reality.

## The 68–95–99.7 Rule

### One in a Million

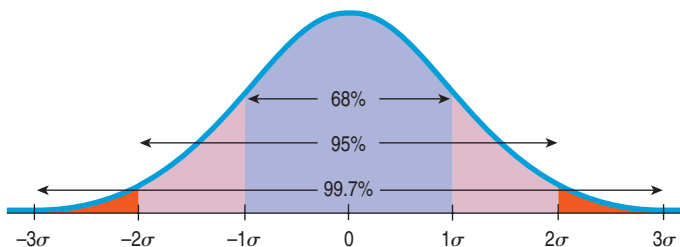
These magic 68, 95, 99.7 values come from the Normal model. As a model, it can give us corresponding values for any  $z$ -score. For example, it tells us that fewer than 1 out of a million values have  $z$ -scores smaller than  $-5.0$  or larger than  $+5.0$ . So if someone tells you you're "one in a million," they must really admire your  $z$ -score.

### TI-*n*spire

**The 68–95–99.7 Rule.** See it work for yourself.

Normal models give us an idea of how extreme a value is by telling us how likely it is to find one that far from the mean. We'll soon show how to find these numbers precisely—but one simple rule is usually all we need.

It turns out that in a Normal model, about 68% of the values fall within 1 standard deviation of the mean, about 95% of the values fall within 2 standard deviations of the mean, and about 99.7%—almost all—of the values fall within 3 standard deviations of the mean. These facts are summarized in a rule that we call (let's see . . .) the **68–95–99.7 Rule**.<sup>4</sup>



**FIGURE 6.6**

Reaching out one, two, and three standard deviations on a Normal model gives the 68–95–99.7 Rule, seen as proportions of the area under the curve.

### FOR EXAMPLE

#### Using the 68–95–99.7 Rule

**Question:** In the 2006 Winter Olympics men's combined event, Jean-Baptiste Grange of France skied the slalom in 88.46 seconds—about 1 standard deviation faster than the mean. If a Normal model is useful in describing slalom times, about how many of the 35 skiers finishing the event would you expect skied the slalom *faster* than Jean-Baptiste?

From the 68–95–99.7 Rule, we expect 68% of the skiers to be within one standard deviation of the mean. Of the remaining 32%, we expect half on the high end and half on the low end. 16% of 35 is 5.6, so, conservatively, we'd expect about 5 skiers to do better than Jean-Baptiste.



### JUST CHECKING

4. As a group, the Dutch are among the tallest people in the world. The average Dutch man is 184 cm tall—just over 6 feet (and the average Dutch woman is 170.8 cm tall—just over 5'7"). If a Normal model is appropriate and the standard deviation for men is about 8 cm, what percentage of all Dutch men will be over 2 meters (6'6") tall?
5. Suppose it takes you 20 minutes, on average, to drive to school, with a standard deviation of 2 minutes. Suppose a Normal model is appropriate for the distributions of driving times.
  - a) How often will you arrive at school in less than 22 minutes?
  - b) How often will it take you more than 24 minutes?
  - c) Do you think the distribution of your driving times is unimodal and symmetric?
  - d) What does this say about the accuracy of your predictions? Explain.

<sup>4</sup> This rule is also called the "Empirical Rule" because it originally came from observation. The rule was first published by Abraham de Moivre in 1733, 75 years before the Normal model was discovered. Maybe it should be called "de Moivre's Rule," but that wouldn't help us remember the important numbers, 68, 95, and 99.7.

## The First Three Rules for Working with Normal Models

**AS** **Activity: Working with Normal Models.** Well, actually playing with them. This interactive tool lets you do what this chapter's figures can't do, move them!

1. Make a picture.
2. Make a picture.
3. Make a picture.

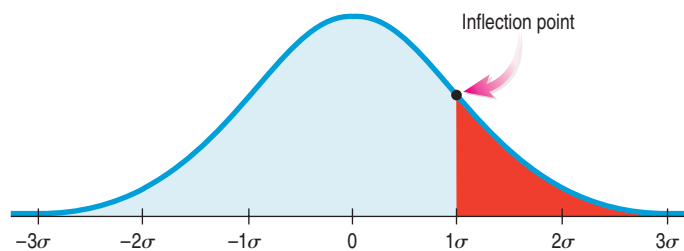
Although we're thinking about models, not histograms of data, the three rules don't change. To help you think clearly, a simple hand-drawn sketch is all you need. Even experienced statisticians sketch pictures to help them think about Normal models. You should too.

Of course, when we have data, we'll also need to make a histogram to check the **Nearly Normal Condition** to be sure we can use the Normal model to model the data's distribution. Other times, we may be told that a Normal model is appropriate based on prior knowledge of the situation or on theoretical considerations.

**AS** **Activity: Normal Models.** Normal models have several interesting properties—see them here.

**How to Sketch a Normal Curve That Looks Normal** To sketch a good Normal curve, you need to remember only three things:

- ▶ The Normal curve is bell-shaped and symmetric around its mean. Start at the middle, and sketch to the right and left from there.
- ▶ Even though the Normal model extends forever on either side, you need to draw it only for 3 standard deviations. After that, there's so little left that it isn't worth sketching.
- ▶ The place where the bell shape changes from curving downward to curving back up—the *inflection point*—is exactly one standard deviation away from the mean.



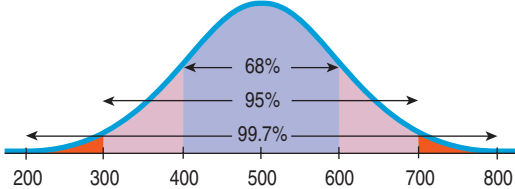
### STEP-BY-STEP EXAMPLE

#### Working with the 68–95–99.7 Rule

The SAT Reasoning Test has three parts: Writing, Math, and Critical Reading (Verbal). Each part has a distribution that is roughly unimodal and symmetric and is designed to have an overall mean of about 500 and a standard deviation of 100 for all test takers. In any one year, the mean and standard deviation may differ from these target values by a small amount, but they are a good overall approximation.

**Question:** Suppose you earned a 600 on one part of your SAT. Where do you stand among all students who took that test?

You could calculate your  $z$ -score and find out that it's  $z = (600 - 500)/100 = 1.0$ , but what does that tell you about your percentile? You'll need the Normal model and the 68–95–99.7 Rule to answer that question.

<p><b>THINK</b></p> <p><b>Plan</b> State what you want to know.</p> <p><b>Variables</b> Identify the variable and report the W's.</p> <p>Be sure to check the appropriate conditions.</p> <p>Specify the parameters of your model.</p>	<p>I want to see how my SAT score compares with the scores of all other students. To do that, I'll need to model the distribution.</p> <p>Let <math>y =</math> my SAT score. Scores are quantitative but have no meaningful units other than points.</p> <p>✓ <b>Nearly Normal Condition:</b> If I had data, I would check the histogram. I have no data, but I am told that the SAT scores are roughly unimodal and symmetric.</p> <p>I will model SAT score with a <math>N(500, 100)</math> model.</p>
<p><b>SHOW</b></p> <p><b>Mechanics</b> Make a picture of this Normal model. (A simple sketch is all you need.)</p> <p>Locate your score.</p>	 <p>My score of 600 is 1 standard deviation above the mean. That corresponds to one of the points of the 68–95–99.7 Rule.</p>
<p><b>TELL</b></p> <p><b>Conclusion</b> Interpret your result in context.</p>	<p>About 68% of those who took the test had scores that fell no more than 1 standard deviation from the mean, so <math>100\% - 68\% = 32\%</math> of all students had scores more than 1 standard deviation away. Only half of those were on the high side, so about 16% (half of 32%) of the test scores were better than mine. My score of 600 is higher than about 84% of all scores on this test.</p>

The bounds of SAT scoring at 200 and 800 can also be explained by the 68–95–99.7 Rule. Since 200 and 800 are three standard deviations from 500, it hardly pays to extend the scoring any farther on either side. We'd get more information only on  $100 - 99.7 = 0.3\%$  of students.

**The Worst-Case Scenario\*** Suppose we encounter an observation that's 5 standard deviations above the mean. Should we be surprised? We've just seen that when a Normal model is appropriate, such a value is exceptionally rare. After all, 99.7% of all the values should be within 3 standard deviations of the mean, so anything farther away would be unusual indeed.

But our handy 68–95–99.7 Rule applies only to Normal models, and the Normal is such a *nice* shape. What if we're dealing with a distribution that's strongly

skewed (like the CEO salaries), or one that is uniform or bimodal or something really strange? A Normal model has 68% of its observations within one standard deviation of the mean, but a bimodal distribution could even be entirely empty in the middle. In that case could we still say anything at all about an observation 5 standard deviations above the mean?

Remarkably, even with really weird distributions, the worst case can't get all that bad. A Russian mathematician named Pafnuty Tchebycheff<sup>5</sup> answered the question by proving this theorem:

*In any distribution, at least  $1 - \frac{1}{k^2}$  of the values must lie within  $\pm k$  standard deviations of the mean.*

What does that mean?

- ▶ For  $k = 1$ ,  $1 - \frac{1}{1^2} = 0$ ; if the distribution is far from Normal, it's possible that none of the values are within 1 standard deviation of the mean. We should be really cautious about saying anything about 68% unless we think a Normal model is justified. (Tchebycheff's theorem really is about the worst case; it tells us nothing about the middle; only about the extremes.)
- ▶ For  $k = 2$ ,  $1 - \frac{1}{2^2} = \frac{3}{4}$ ; no matter how strange the shape of the distribution, at least 75% of the values must be within 2 standard deviations of the mean. Normal models may expect 95% in that 2-standard-deviation interval, but even in a worst-case scenario it can never go lower than 75%.
- ▶ For  $k = 3$ ,  $1 - \frac{1}{3^2} = \frac{8}{9}$ ; in any distribution, at least 89% of the values lie within 3 standard deviations of the mean.

What we see is that values beyond 3 standard deviations from the mean are uncommon, Normal model or not. Tchebycheff tells us that at least 96% of all values must be within 5 standard deviations of the mean. While we can't always apply the 68–95–99.7 Rule, we can be sure that the observation we encountered 5 standard deviations above the mean is unusual.

## Finding Normal Percentiles

**AS** **Activity: Your Pulse z-Score.** Is your pulse rate high or low? Find its z-score with the Normal Model Tool.

An SAT score of 600 is easy to assess, because we can think of it as one standard deviation above the mean. If your score was 680, though, where do you stand among the rest of the people tested? Your z-score is 1.80, so you're somewhere between 1 and 2 standard deviations above the mean. We figured out that no more than 16% of people score better than 600. By the same logic, no more than 2.5% of people score better than 700. Can we be more specific than "between 16% and 2.5%"?

When the value doesn't fall exactly 1, 2, or 3 standard deviations from the mean, we can look it up in a table of **Normal percentiles** or use technology.<sup>6</sup> Either way, we first convert our data to z-scores before using the table. Your SAT score of 680 has a z-score of  $(680 - 500)/100 = 1.80$ .

**AS** **Activity: The Normal Table.** Table Z just sits there, but this version of the Normal table changes so it always Makes a Picture that fits.

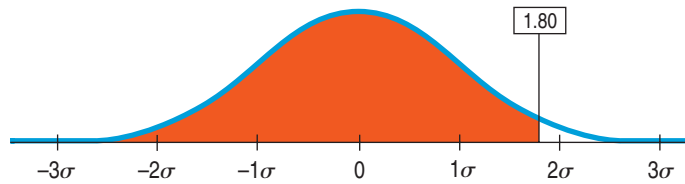
<sup>5</sup> He may have made the worst case for deviations clear, but the English spelling of his name is not. You'll find his first name spelled Pavnutii or Pavnuty and his last name spelled Chebsheff, Cebysev, and other creative versions.

<sup>6</sup> See Table Z in Appendix G, if you're curious. But your calculator (and any statistics computer package) does this, too—and more easily!



**FIGURE 6.7**

A table of Normal percentiles (Table Z in Appendix G) lets us find the percentage of individuals in a Standard Normal distribution falling below any specified z-score value.



z	.00	.01
1.7	.9554	.9564
1.8	.9641	.9649
1.9	.9713	.9719

**TI-*nspire***  
**Normal percentiles.** Explore the relationship between z-scores and areas in a Normal model.

In the piece of the table shown, we find your z-score by looking down the left column for the first two digits, 1.8, and across the top row for the third digit, 0. The table gives the percentile as 0.9641. That means that 96.4% of the z-scores are less than 1.80. Only 3.6% of people, then, scored better than 680 on the SAT.

Most of the time, though, you'll do this with your calculator.

**TI Tips** **Finding Normal percentages**

```
Plot1 Plot2 Plot3
Y1=normalpdf(X)
V2=
V3=
V4=
V5=
V6=
```

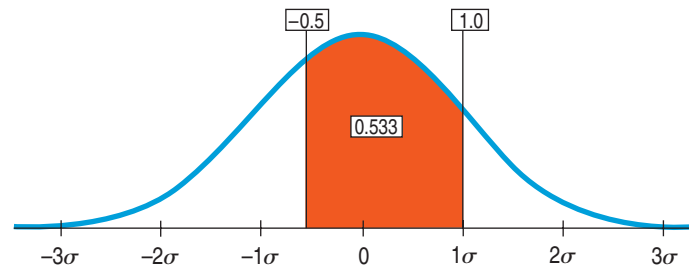


Your calculator knows the Normal model. Have a look under **2nd DISTR**. There you will see three “norm” functions, **normalpdf()**, **normalcdf()**, and **invNorm()**. Let’s play with the first two.

- **normalpdf()** calculates y-values for graphing a Normal curve. You probably won’t use this very often, if at all. If you want to try it, graph **Y1=normalpdf(X)** in a graphing WINDOW with **Xmin=-4**, **Xmax=4**, **Ymin=-0.1**, and **Ymax=0.5**.
- **normalcdf()** finds the proportion of area under the curve between two z-score cut points, by specifying **normalcdf(zLeft, zRight)**. Do make friends with this function; you will use it often!

**Example 1**

The Normal model shown shades the region between  $z = -0.5$  and  $z = 1.0$ .



To find the shaded area:

Under **2nd DISTR** select **normalcdf()**; hit **ENTER**.

Specify the cut points: **normalcdf(-.5, 1.0)** and hit **ENTER** again.

There’s the area. Approximately 53% of a Normal model lies between half a standard deviation below and one standard deviation above the mean.

**Example 2**

In the example in the text we used Table Z to determine the fraction of SAT scores above your score of 680. Now let’s do it again, this time using your TI.

First we need z-scores for the cut points:

- Since 680 is 1.8 standard deviations above the mean, your z-score is 1.8; that’s the left cut point.

```
normalcdf(-.5, 1.
0)
.5328072002
```

```
normalcdf(1.8,99)
.0359302655
```

- Theoretically the standard Normal model extends rightward forever, but you can't tell the calculator to use infinity as the right cut point. Recall that for a Normal model almost all the area lies within  $\pm 3$  standard deviations of the mean, so any upper cut point beyond, say,  $z = 5$  does not cut off anything very important. We suggest you always use 99 (or  $-99$ ) when you really want infinity as your cut point—it's easy to remember and way beyond any meaningful area.

Now you're ready. Use the command `normalcdf(1.8,99)`.

There you are! The Normal model estimates that approximately 3.6% of SAT scores are higher than 680.

### STEP-BY-STEP EXAMPLE

### Working with Normal Models Part I

The Normal model is our first model for data. It's the first in a series of modeling situations where we step away from the data at hand to make more general statements about the world. We'll become more practiced in thinking about and learning the details of models as we progress through the book. To give you some practice in thinking about the Normal model, here are several problems that ask you to find percentiles in detail.

**Question:** What proportion of SAT scores fall between 450 and 600?

THINK

**Plan** State the problem.

**Variables** Name the variable.

Check the appropriate conditions and specify which Normal model to use.

I want to know the proportion of SAT scores between 450 and 600.

Let  $y = \text{SAT score}$ .

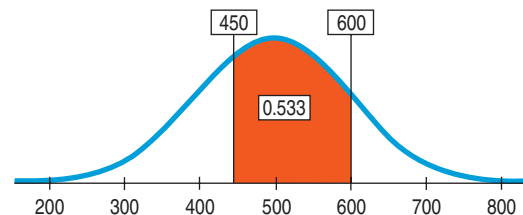
✓ **Nearly Normal Condition:** We are told that SAT scores are nearly Normal.

I'll model SAT scores with a  $N(500, 100)$  model, using the mean and standard deviation specified for them.

SHOW

**Mechanics** Make a picture of this Normal model. Locate the desired values and shade the region of interest.

Find  $z$ -scores for the cut points 450 and 600. Use technology to find the desired proportions, represented by the area under the curve. (This was Example 1 in the TI Tips—take another look.)



Standardizing the two scores, I find that

$$z = \frac{(y - \mu)}{\sigma} = \frac{(600 - 500)}{100} = 1.00$$

and

$$z = \frac{(450 - 500)}{100} = -0.50$$

(If you use a table, then you need to subtract the two areas to find the area *between* the cut points.)

So,

$$\begin{aligned} \text{Area}(450 < y < 600) &= \text{Area}(-0.5 < z < 1.0) \\ &= 0.5328 \end{aligned}$$

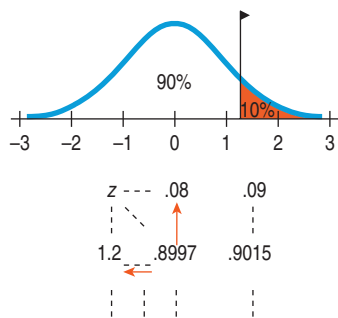
(OR: From Table Z, the area ( $z < 1.0$ ) = 0.8413 and area ( $z < -0.5$ ) = 0.3085, so the proportion of z-scores *between* them is  $0.8413 - 0.3085 = 0.5328$ , or 53.28%.)



**Conclusion** Interpret your result in context.

The Normal model estimates that about 53.3% of SAT scores fall between 450 and 600.

## From Percentiles to Scores: z in Reverse



Finding areas from z-scores is the simplest way to work with the Normal model. But sometimes we start with areas and are asked to work backward to find the corresponding z-score or even the original data value. For instance, what z-score cuts off the top 10% in a Normal model?

Make a picture like the one shown, shading the rightmost 10% of the area. Notice that this is the 90th percentile. Look in Table Z for an area of 0.900. The exact area is not there, but 0.8997 is pretty close. That shows up in the table with 1.2 in the left margin and .08 in the top margin. The z-score for the 90th percentile, then, is approximately  $z = 1.28$ .

Computers and calculators will determine the cut point more precisely (and more easily).

### TI Tips

```
invNorm(.25)
-.6744897495
```

```
invNorm(.9)
1.281551567
```

### Finding Normal cutpoints

To find the z-score at the 25th percentile, go to 2nd DISTR again. This time we'll use the third of the "norm" functions, `invNorm(.`

Just specify the desired percentile with the command `invNorm(.25)` and hit ENTER. The calculator says that the cut point for the leftmost 25% of a Normal model is approximately  $z = -0.674$ .

One more example: What z-score cuts off the highest 10% of a Normal model? That's easily done—just remember to specify the *percentile*. Since we want the cut point for the *highest* 10%, we know that the other 90% must be *below* that z-score. The cut point, then, must stand at the 90th percentile, so specify `invNorm(.90)`.

Only 10% of the area in a Normal model is more than about 1.28 standard deviations above the mean.

## STEP-BY-STEP EXAMPLE

## Working with Normal Models Part II

**Question:** Suppose a college says it admits only people with SAT Verbal test scores among the top 10%. How high a score does it take to be eligible?

THINK

**Plan** State the problem.

**Variable** Define the variable.

Check to see if a Normal model is appropriate, and specify which Normal model to use.

How high an SAT Verbal score do I need to be in the top 10% of all test takers?

Let  $y$  = my SAT score.

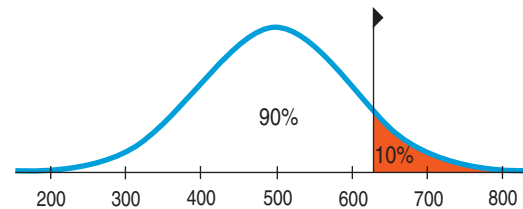
✓ **Nearly Normal Condition:** I am told that SAT scores are nearly Normal. I'll model them with  $N(500, 100)$ .

SHOW

**Mechanics** Make a picture of this Normal model. Locate the desired percentile approximately by shading the rightmost 10% of the area.

The college takes the top 10%, so its cutoff score is the 90th percentile. Find the corresponding  $z$ -score using your calculator as shown in the TI Tips. (OR: Use Table Z as shown on p. 119.)

Convert the  $z$ -score back to the original units.



The cut point is  $z = 1.28$ .

A  $z$ -score of 1.28 is 1.28 standard deviations above the mean. Since the SD is 100, that's 128 SAT points. The cutoff is 128 points above the mean of 500, or 628.

TELL

**Conclusion** Interpret your results in the proper context.

Because the school wants SAT Verbal scores in the top 10%, the cutoff is 628. (Actually, since SAT scores are reported only in multiples of 10, I'd have to score at least a 630.)

TI-*n*spire

**Normal models.** Watch the Normal model react as you change the mean and standard deviation.

## STEP-BY-STEP EXAMPLE

## More Working with Normal Models

Working with Normal percentiles can be a little tricky, depending on how the problem is stated. Here are a few more worked examples of the kind you're likely to see.

*A cereal manufacturer has a machine that fills the boxes. Boxes are labeled "16 ounces," so the company wants to have that much cereal in each box, but since no packaging process is perfect, there will be minor variations. If the machine is set at exactly 16 ounces and the Normal model applies (or at least the distribution is roughly symmetric), then about half of the boxes will be underweight, making consumers unhappy and exposing the company to bad publicity and possible lawsuits. To prevent underweight boxes, the manufacturer has to set the mean a little higher than 16.0 ounces.*

*Based on their experience with the packaging machine, the company believes that the amount of cereal in the boxes fits a Normal model with a standard deviation of 0.2 ounces. The manufacturer decides to set the machine to put an average of 16.3 ounces in each box. Let's use that model to answer a series of questions about these cereal boxes.*

**Question 1:** What fraction of the boxes will be underweight?

THINK

**Plan** State the problem.

**Variable** Name the variable.

Check to see if a Normal model is appropriate.

Specify which Normal model to use.

What proportion of boxes weigh less than 16 ounces?

Let  $y$  = weight of cereal in a box.

✓ **Nearly Normal Condition:** I have no data, so I cannot make a histogram, but I am told that the company believes the distribution of weights from the machine is Normal.

I'll use a  $N(16.3, 0.2)$  model.

SHOW

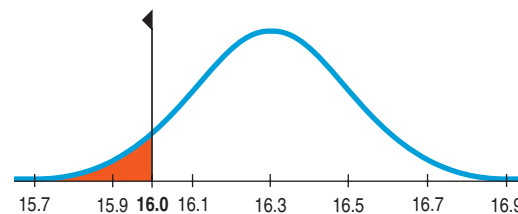
**Mechanics** Make a picture of this Normal model. Locate the value you're interested in on the picture, label it, and shade the appropriate region.

REALITY CHECK

Estimate from the picture the percentage of boxes that are underweight. (This will be useful later to check that your answer makes sense.) It looks like a low percentage. Less than 20% for sure.

Convert your cutoff value into a z-score.

Find the area with your calculator (or use the Normal table).



I want to know what fraction of the boxes will weigh less than 16 ounces.

$$z = \frac{y - \mu}{\sigma} = \frac{16 - 16.3}{0.2} = -1.50$$

$$\text{Area}(y < 16) = \text{Area}(z < -1.50) = 0.0668$$



**Conclusion** State your conclusion, and check that it's consistent with your earlier guess. It's below 20%—seems okay.

I estimate that approximately 6.7% of the boxes will contain less than 16 ounces of cereal.

**Question 2:** The company's lawyers say that 6.7% is too high. They insist that no more than 4% of the boxes can be underweight. So the company needs to set the machine to put a little more cereal in each box. What mean setting do they need?



**Plan** State the problem.

**Variable** Name the variable.

Check to see if a Normal model is appropriate.

Specify which Normal model to use. This time you are not given a value for the mean!



We found out earlier that setting the machine to  $\mu = 16.3$  ounces made 6.7% of the boxes too light. We'll need to raise the mean a bit to reduce this fraction.

What mean weight will reduce the proportion of underweight boxes to 4%?

Let  $y$  = weight of cereal in a box.

✓ **Nearly Normal Condition:** I am told that a Normal model applies.

I don't know  $\mu$ , the mean amount of cereal. The standard deviation for this machine is 0.2 ounces. The model is  $N(\mu, 0.2)$ .

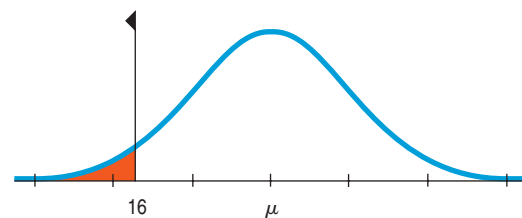
No more than 4% of the boxes can be below 16 ounces.



**Mechanics** Make a picture of this Normal model. Center it at  $\mu$  (since you don't know the mean), and shade the region below 16 ounces.

Using your calculator (or the Normal table), find the  $z$ -score that cuts off the lowest 4%.

Use this information to find  $\mu$ . It's located 1.75 standard deviations to the right of 16. Since  $\sigma$  is 0.2, that's  $1.75 \times 0.2$ , or 0.35 ounces more than 16.



The  $z$ -score that has 0.04 area to the left of it is  $z = -1.75$ .

For 16 to be 1.75 standard deviations below the mean, the mean must be

$$16 + 1.75(0.2) = 16.35 \text{ ounces.}$$



**Conclusion** Interpret your result in context. (This makes sense; we knew it would have to be just a bit higher than 16.3.)

The company must set the machine to average 16.35 ounces of cereal per box.

**Question 3:** The company president vetoes that plan, saying the company should give away less free cereal, not more. Her goal is to set the machine no higher than 16.2 ounces and still have only 4% underweight boxes. The only way to accomplish this is to reduce the standard deviation. What standard deviation must the company achieve, and what does that mean about the machine?

**THINK**

**Plan** State the problem.

**Variable** Name the variable.

Check conditions to be sure that a Normal model is appropriate.

Specify which Normal model to use. This time you don't know  $\sigma$ .

**REALITY CHECK**

We know the new standard deviation must be less than 0.2 ounces.

What standard deviation will allow the mean to be 16.2 ounces and still have only 4% of boxes underweight?

Let  $y$  = weight of cereal in a box.

✓ **Nearly Normal Condition:** The company believes that the weights are described by a Normal model.

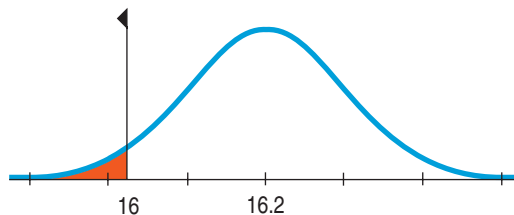
I know the mean, but not the standard deviation, so my model is  $N(16.2, \sigma)$ .

**SHOW**

**Mechanics** Make a picture of this Normal model. Center it at 16.2, and shade the area you're interested in. We want 4% of the area to the left of 16 ounces.

Find the z-score that cuts off the lowest 4%.

Solve for  $\sigma$ . (We need 16 to be 1.75  $\sigma$ 's below 16.2, so  $1.75\sigma$  must be 0.2 ounces. You could just start with that equation.)



I know that the z-score with 4% below it is  $z = -1.75$ .

$$z = \frac{y - \mu}{\sigma}$$

$$-1.75 = \frac{16 - 16.2}{\sigma}$$

$$1.75\sigma = 0.2$$

$$\sigma = 0.114$$

**TELL**

**Conclusion** Interpret your result in context.

As we expected, the standard deviation is lower than before—actually, quite a bit lower.

The company must get the machine to box cereal with a standard deviation of only 0.114 ounces. This means the machine must be more consistent (by nearly a factor of 2) in filling the boxes.

## Are You Normal? Find Out with a Normal Probability Plot

### TI-*n*spire

**Normal probability plots and histograms.** See how a normal probability plot responds as you change the shape of a distribution.

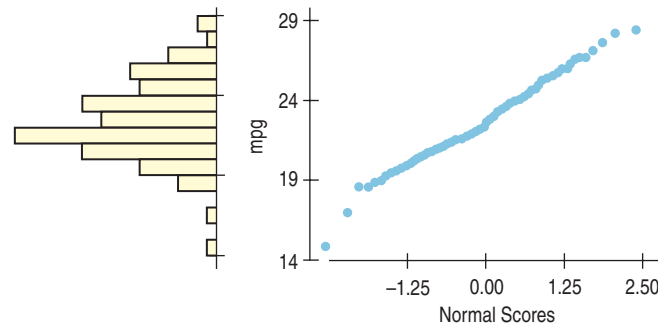
In the examples we've worked through, we've assumed that the underlying data distribution was roughly unimodal and symmetric, so that using a Normal model makes sense. When you have data, you must *check* to see whether a Normal model is reasonable. How? Make a picture, of course! Drawing a histogram of the data and looking at the shape is one good way to see if a Normal model might work.

There's a more specialized graphical display that can help you to decide whether the Normal model is appropriate: the **Normal probability plot**. If the distribution of the data is roughly Normal, the plot is roughly a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal. This plot is usually able to show deviations from Normality more clearly than the corresponding histogram, but it's usually easier to understand *how* a distribution fails to be Normal by looking at its histogram.

Some data on a car's fuel efficiency provide an example of data that are nearly Normal. The overall pattern of the Normal probability plot is straight. The two trailing low values correspond to the values in the histogram that trail off the low end. They're not quite in line with the rest of the data set. The Normal probability plot shows us that they're a bit lower than we'd expect of the lowest two values in a Normal model.

**FIGURE 6.9**

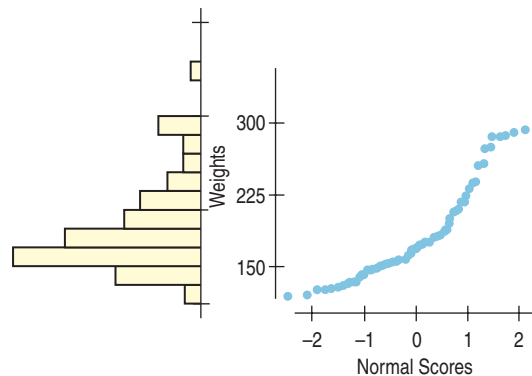
*Histogram and Normal probability plot for gas mileage (mpg) recorded by one of the authors over the 8 years he owned a 1989 Nissan Maxima. The vertical axes are the same, so each dot on the probability plot would fall into the bar on the histogram immediately to its left.*



By contrast, the Normal probability plot of the men's *Weights* from the NHANES Study is far from straight. The weights are skewed to the high end, and the plot is curved. We'd conclude from these pictures that approximations using the 68–95–99.7 Rule for these data would not be very accurate.

**FIGURE 6.10**

*Histogram and Normal probability plot for men's weights. Note how a skewed distribution corresponds to a bent probability plot.*





## TI Tips

## Creating a Normal probability plot

Let's make a Normal probability plot with the calculator. Here are the boys' agility test scores we looked at in Chapter 5; enter them in **L1**:

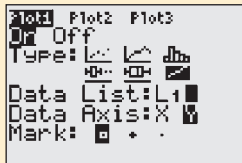
22, 17, 18, 29, 22, 23, 24, 23, 17, 21

Now you can create the plot:

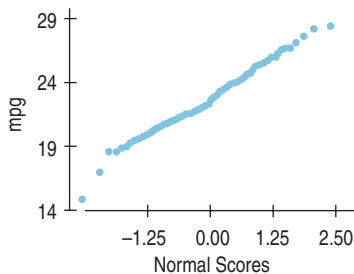
- Turn a **STATPLOT** On.
- Tell it to make a Normal probability plot by choosing the last of the icons.
- Specify your datalist and which axis you want the data on. (We'll use **Y** so the plot looks like the others we showed you.)
- Specify the **Mark** you want the plot to use.
- Now **ZoomStat** does the rest.

The plot doesn't look very straight. Normality is certainly questionable here.

(Not that it matters in making this decision, but that vertical line is the  $y$ -axis. Points to the left have negative  $z$ -scores and points to the right have positive  $z$ -scores.)



## How Does a Normal Probability Plot Work?



**A S** **Activity: Assessing Normality.** This activity guides you through the process of checking the Nearly Normal condition using your statistics package.

Why does the Normal probability plot work like that? We looked at 100 fuel efficiency measures for the author's Nissan car. The smallest of these has a  $z$ -score of  $-3.16$ . The Normal model can tell us what value to expect for the smallest  $z$ -score in a batch of 100 if a Normal model were appropriate. That turns out to be  $-2.58$ . So our first data value is smaller than we would expect from the Normal.

We can continue this and ask a similar question for each value. For example, the 14th-smallest fuel efficiency has a  $z$ -score of almost exactly  $-1$ , and that's just what we should expect (well,  $-1.1$  to be exact). A Normal probability plot takes each data value and plots it against the  $z$ -score you'd expect that point to have if the distribution were perfectly Normal.<sup>7</sup>

When the values match up well, the line is straight. If one or two points are surprising from the Normal's point of view, they don't line up. When the entire distribution is skewed or different from the Normal in some other way, the values don't match up very well at all and the plot bends.

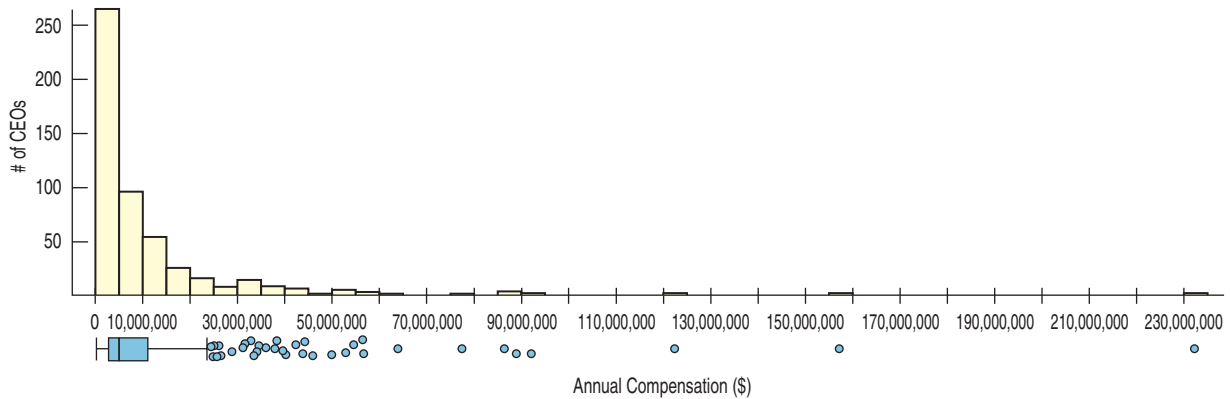
It turns out to be tricky to find the values we expect. They're called *Normal scores*, but you can't easily look them up in the tables. That's why probability plots are best made with technology and not by hand.

The best advice on using Normal probability plots is to see whether they are straight. If so, then your data look like data from a Normal model. If not, make a histogram to understand how they differ from the model.

<sup>7</sup> Sometimes the Normal probability plot switches the two axes, putting the data on the  $x$ -axis and the  $z$ -scores on the  $y$ -axis.

## WHAT CAN GO WRONG?

- ▶ **Don't use a Normal model when the distribution is not unimodal and symmetric.** Normal models are so easy and useful that it is tempting to use them even when they don't describe the data very well. That can lead to wrong conclusions. Don't use a Normal model without first checking the **Nearly Normal Condition**. Look at a picture of the data to check that it is unimodal and symmetric. A histogram, or a Normal probability plot, can help you tell whether a Normal model is appropriate.



The CEOs (p. 90) had a mean total compensation of \$10,307,311.87 with a standard deviation of \$17,964,615.16. Using the Normal model rule, we should expect about 68% of the CEOs to have compensations between  $-\$7,657,303.29$  and  $\$28,271,927.03$ . In fact, more than 90% of the CEOs have annual compensations in this range. What went wrong? The distribution is skewed, not symmetric. Using the 68–95–99.7 Rule for data like these will lead to silly results.

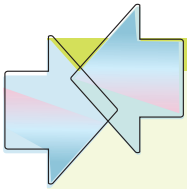
- ▶ **Don't use the mean and standard deviation when outliers are present.** Both means and standard deviations can be distorted by outliers, and no model based on distorted values will do a good job. A z-score calculated from a distribution with outliers may be misleading. It's always a good idea to check for outliers. How? Make a picture.
- ▶ **Don't round your results in the middle of a calculation.** We reported the mean of the heptathletes' long jump as 6.16 meters. More precisely, it was 6.16153846153846 meters.

You should use all the precision available in the data for all the intermediate steps of a calculation. Using the more precise value for the mean (and also carrying 15 digits for the SD), the z-score calculation for Klüff's long jump comes out to

$$z = \frac{6.78 - 6.16153846153846}{0.2297597407326585} = 2.691775053755667700$$

We'd report that as 2.692, as opposed to the rounded-off value of 2.70 we got earlier from the table.

- ▶ **Don't worry about minor differences in results.** Because various calculators and programs may carry different precision in calculations, your answers may differ slightly from those we show in the text and in the Step-By-Steps, or even from the values given in the answers in the back of the book. Those differences aren't anything to worry about. They're not the main story Statistics tries to tell.



## CONNECTIONS

Changing the center and spread of a variable is equivalent to changing its *units*. Indeed, the only part of the data's context changed by standardizing is the units. All other aspects of the context do not depend on the choice or modification of measurement units. This fact points out an important distinction between the numbers the data provide for calculation and the meaning of the variables and the relationships among them. Standardizing can make the numbers easier to work with, but it does not alter the meaning.

Another way to look at this is to note that standardizing may change the center and spread values, but it does not affect the *shape* of a distribution. A histogram or boxplot of standardized values looks just the same as the histogram or boxplot of the original values except, perhaps, for the numbers on the axes.

When we summarized *shape*, *center*, and *spread* for histograms, we compared them to unimodal, symmetric shapes. You couldn't ask for a nicer example than the Normal model. And if the shape *is* like a Normal, we'll use the the mean and standard deviation to standardize the values.



## WHAT HAVE WE LEARNED?

We've learned that the story data can tell may be easier to understand after shifting or rescaling the data.

- ▶ Shifting data by adding or subtracting the same amount from each value affects measures of center and position but not measures of spread.
- ▶ Rescaling data by multiplying or dividing every value by a constant, changes all the summary statistics—center, position, and spread.

We've learned the power of standardizing data.

- ▶ Standardizing uses the standard deviation as a ruler to measure distance from the mean, creating z-scores.
- ▶ Using these z-scores, we can compare apples and oranges—values from different distributions or values based on different units.
- ▶ And a z-score can identify unusual or surprising values among data.

We've learned that the 68–95–99.7 Rule can be a useful rule of thumb for understanding distributions.

- ▶ For data that are unimodal and symmetric, about 68% fall within 1 SD of the mean, 95% fall within 2 SDs of the mean, and 99.7% fall within 3 SDs of the mean (see p. 130).

Again we've seen the importance of *Thinking* about whether a method will work.

- ▶ **Normality Assumption:** We sometimes work with Normal tables (Table Z). Those tables are based on the Normal model.
- ▶ Data can't be exactly Normal, so we check the **Nearly Normal Condition** by making a histogram (is it unimodal, symmetric, and free of outliers?) or a Normal probability plot (is it straight enough?). (See p. 125.)

## Terms

**Standardizing**

105. We standardize to eliminate units. Standardized values can be compared and combined even if the original variables had different units and magnitudes.

**Standardized value**

105. A value found by subtracting the mean and dividing by the standard deviation.

Shifting	107. Adding a constant to each data value adds the same constant to the mean, the median, and the quartiles, but does not change the standard deviation or IQR.
Rescaling	108. Multiplying each data value by a constant multiplies both the measures of position (mean, median, and quartiles) and the measures of spread (standard deviation and IQR) by that constant.
Normal model	112. A useful family of models for unimodal, symmetric distributions.
Parameter	112. A numerically valued attribute of a model. For example, the values of $\mu$ and $\sigma$ in a $N(\mu, \sigma)$ model are parameters.
Statistic	112. A value calculated from data to summarize aspects of the data. For example, the mean, $\bar{y}$ and standard deviation, $s$ , are statistics.
z-score	105. A z-score tells how many standard deviations a value is from the mean; z-scores have a mean of 0 and a standard deviation of 1. When working with data, use the statistics $\bar{y}$ and $s$ : $z = \frac{y - \bar{y}}{s}.$
	112. When working with models, use the parameters $\mu$ and $\sigma$ : $z = \frac{y - \mu}{\sigma}.$
Standard Normal model	112. A Normal model, $N(\mu, \sigma)$ with mean $\mu = 0$ and standard deviation $\sigma = 1$ . Also called the <b>standard Normal distribution</b> .
Nearly Normal Condition	112. A distribution is nearly Normal if it is unimodal and symmetric. We can check by looking at a histogram or a Normal probability plot.
68–95–99.7 Rule	113. In a Normal model, about 68% of values fall within 1 standard deviation of the mean, about 95% fall within 2 standard deviations of the mean, and about 99.7% fall within 3 standard deviations of the mean.
Normal percentile	116. The Normal percentile corresponding to a z-score gives the percentage of values in a standard Normal distribution found at that z-score or below.
Normal probability plot	124. A display to help assess whether a distribution of data is approximately Normal. If the plot is nearly straight, the data satisfy the <b>Nearly Normal Condition</b> .

## Skills



- ▶ Understand how adding (subtracting) a constant or multiplying (dividing) by a constant changes the center and/or spread of a variable.
- ▶ Recognize when standardization can be used to compare values.
- ▶ Understand that standardizing uses the standard deviation as a ruler.
- ▶ Recognize when a Normal model is appropriate.



- ▶ Know how to calculate the z-score of an observation.
- ▶ Know how to compare values of two different variables using their z-scores.
- ▶ Be able to use Normal models and the 68–95–99.7 Rule to estimate the percentage of observations falling within 1, 2, or 3 standard deviations of the mean.
- ▶ Know how to find the percentage of observations falling below any value in a Normal model using a Normal table or appropriate technology.
- ▶ Know how to check whether a variable satisfies the **Nearly Normal Condition** by making a Normal probability plot or a histogram.



- ▶ Know what z-scores mean.
- ▶ Be able to explain how extraordinary a standardized value may be by using a Normal model.

## NORMAL PLOTS ON THE COMPUTER

The best way to tell whether your data can be modeled well by a Normal model is to make a picture or two. We've already talked about making histograms. Normal probability plots are almost never made by hand because the values of the Normal scores are tricky to find. But most statistics software make Normal plots, though various packages call the same plot by different names and array the information differently.

## EXERCISES

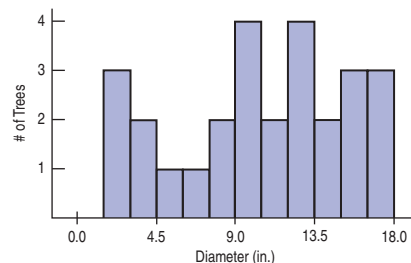
1. **Shipments.** A company selling clothing on the Internet reports that the packages it ships have a median weight of 68 ounces and an IQR of 40 ounces.
  - a) The company plans to include a sales flyer weighing 4 ounces in each package. What will the new median and IQR be?
  - b) If the company recorded the shipping weights of these new packages in pounds instead of ounces, what would the median and IQR be? (1 lb. = 16 oz.)
2. **Hotline.** A company's customer service hotline handles many calls relating to orders, refunds, and other issues. The company's records indicate that the median length of calls to the hotline is 4.4 minutes with an IQR of 2.3 minutes.
  - a) If the company were to describe the duration of these calls in seconds instead of minutes, what would the median and IQR be?
  - b) In an effort to speed up the customer service process, the company decides to streamline the series of push-button menus customers must navigate, cutting the time by 24 seconds. What will the median and IQR of the length of hotline calls become?
3. **Payroll.** Here are the summary statistics for the weekly payroll of a small company: lowest salary = \$300, mean salary = \$700, median = \$500, range = \$1200, IQR = \$600, first quartile = \$350, standard deviation = \$400.
  - a) Do you think the distribution of salaries is symmetric, skewed to the left, or skewed to the right? Explain why.
  - b) Between what two values are the middle 50% of the salaries found?
  - c) Suppose business has been good and the company gives every employee a \$50 raise. Tell the new value of each of the summary statistics.
  - d) Instead, suppose the company gives each employee a 10% raise. Tell the new value of each of the summary statistics.
4. **Hams.** A specialty foods company sells "gourmet hams" by mail order. The hams vary in size from 4.15 to 7.45 pounds, with a mean weight of 6 pounds and standard deviation of 0.65 pounds. The quartiles and median weights are 5.6, 6.2, and 6.55 pounds.
  - a) Find the range and the IQR of the weights.
  - b) Do you think the distribution of the weights is symmetric or skewed? If skewed, which way? Why?
  - c) If these weights were expressed in ounces (1 pound = 16 ounces) what would the mean, standard deviation, quartiles, median, IQR, and range be?
  - d) When the company ships these hams, the box and packing materials add 30 ounces. What are the mean, standard deviation, quartiles, median, IQR, and range of weights of boxes shipped (in ounces)?
  - e) One customer made a special order of a 10-pound ham. Which of the summary statistics of part d might *not* change if that data value were added to the distribution?
5. **SAT or ACT?** Each year thousands of high school students take either the SAT or the ACT, standardized tests used in the college admissions process. Combined SAT Math and Verbal scores go as high as 1600, while the maximum ACT composite score is 36. Since the two exams use very different scales, comparisons of performance are difficult. A convenient rule of thumb is  $SAT = 40 \times ACT + 150$ ; that is, multiply an ACT score by 40 and add 150 points to estimate the equivalent SAT score. An admissions officer reported the following statistics about the ACT scores of 2355 students who applied to her college one year. Find the summaries of equivalent SAT scores.
 

Lowest score = 19    Mean = 27    Standard deviation = 3  
 Q3 = 30            Median = 28    IQR = 6
6. **Cold U?** A high school senior uses the Internet to get information on February temperatures in the town where he'll be going to college. He finds a Web site with some statistics, but they are given in degrees Celsius. The conversion formula is  $^{\circ}F = 9/5 ^{\circ}C + 32$ . Determine the Fahrenheit equivalents for the summary information below.
 

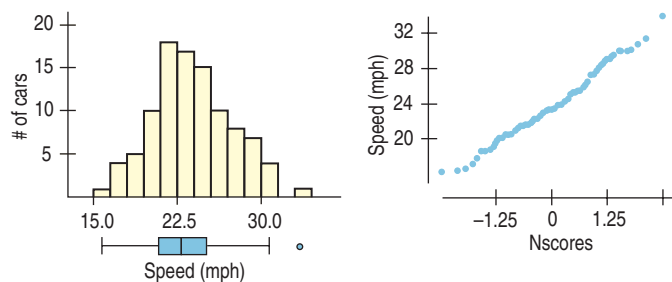
Maximum temperature =  $11^{\circ}C$     Range =  $33^{\circ}$   
 Mean =  $1^{\circ}$     Standard deviation =  $7^{\circ}$   
 Median =  $2^{\circ}$     IQR =  $16^{\circ}$
7. **Stats test.** Suppose your Statistics professor reports test grades as z-scores, and you got a score of 2.20 on an exam. Write a sentence explaining what that means.
8. **Checkup.** One of the authors has an adopted grandson whose birth family members are very short. After examining him at his 2-year checkup, the boy's pediatrician said that the z-score for his height relative to American 2-year-olds was  $-1.88$ . Write a sentence explaining what that means.

9. **Stats test, part II.** The mean score on the Stats exam was 75 points with a standard deviation of 5 points, and Gregor's  $z$ -score was  $-2$ . How many points did he score?
10. **Mensa.** People with  $z$ -scores above 2.5 on an IQ test are sometimes classified as geniuses. If IQ scores have a mean of 100 and a standard deviation of 16 points, what IQ score do you need to be considered a genius?
11. **Temperatures.** A town's January high temperatures average  $36^\circ\text{F}$  with a standard deviation of  $10^\circ$ , while in July the mean high temperature is  $74^\circ$  and the standard deviation is  $8^\circ$ . In which month is it more unusual to have a day with a high temperature of  $55^\circ$ ? Explain.
12. **Placement exams.** An incoming freshman took her college's placement exams in French and mathematics. In French, she scored 82 and in math 86. The overall results on the French exam had a mean of 72 and a standard deviation of 8, while the mean math score was 68, with a standard deviation of 12. On which exam did she do better compared with the other freshmen?
13. **Combining test scores.** The first Stats exam had a mean of 65 and a standard deviation of 10 points; the second had a mean of 80 and a standard deviation of 5 points. Derrick scored an 80 on both tests. Julie scored a 70 on the first test and a 90 on the second. They both totaled 160 points on the two exams, but Julie claims that her total is better. Explain.
14. **Combining scores again.** The first Stat exam had a mean of 80 and a standard deviation of 4 points; the second had a mean of 70 and a standard deviation of 15 points. Reginald scored an 80 on the first test and an 85 on the second. Sara scored an 88 on the first but only a 65 on the second. Although Reginald's total score is higher, Sara feels she should get the higher grade. Explain her point of view.
15. **Final exams.** Anna, a language major, took final exams in both French and Spanish and scored 83 on each. Her roommate Megan, also taking both courses, scored 77 on the French exam and 95 on the Spanish exam. Overall, student scores on the French exam had a mean of 81 and a standard deviation of 5, and the Spanish scores had a mean of 74 and a standard deviation of 15.
- To qualify for language honors, a major must maintain at least an 85 average for all language courses taken. So far, which student qualifies?
  - Which student's overall performance was better?
16. **MP3s.** Two companies market new batteries targeted at owners of personal music players. DuraTunes claims a mean battery life of 11 hours, while RockReady advertises 12 hours.
- Explain why you would also like to know the standard deviations of the battery lifespans before deciding which brand to buy.
  - Suppose those standard deviations are 2 hours for DuraTunes and 1.5 hours for RockReady. You are headed for 8 hours at the beach. Which battery is most likely to last all day? Explain.
  - If your beach trip is all weekend, and you probably will have the music on for 16 hours, which battery is most likely to last? Explain.
17. **Cattle.** The Virginia Cooperative Extension reports that the mean weight of yearling Angus steers is 1152 pounds. Suppose that weights of all such animals can be described by a Normal model with a standard deviation of 84 pounds.
- How many standard deviations from the mean would a steer weighing 1000 pounds be?
  - Which would be more unusual, a steer weighing 1000 pounds or one weighing 1250 pounds?
- T** 18. **Car speeds.** John Beale of Stanford, CA, recorded the speeds of cars driving past his house, where the speed limit read 20 mph. The mean of 100 readings was 23.84 mph, with a standard deviation of 3.56 mph. (He actually recorded every car for a two-month period. These are 100 representative readings.)
- How many standard deviations from the mean would a car going under the speed limit be?
  - Which would be more unusual, a car traveling 34 mph or one going 10 mph?
19. **More cattle.** Recall that the beef cattle described in Exercise 17 had a mean weight of 1152 pounds, with a standard deviation of 84 pounds.
- Cattle buyers hope that yearling Angus steers will weigh at least 1000 pounds. To see how much over (or under) that goal the cattle are, we could subtract 1000 pounds from all the weights. What would the new mean and standard deviation be?
  - Suppose such cattle sell at auction for 40 cents a pound. Find the mean and standard deviation of the sale prices for all the steers.
- T** 20. **Car speeds again.** For the car speed data of Exercise 18, recall that the mean speed recorded was 23.84 mph, with a standard deviation of 3.56 mph. To see how many cars are speeding, John subtracts 20 mph from all speeds.
- What is the mean speed now? What is the new standard deviation?
  - His friend in Berlin wants to study the speeds, so John converts all the original miles-per-hour readings to kilometers per hour by multiplying all speeds by 1.609 (km per mile). What is the mean now? What is the new standard deviation?
21. **Cattle, part III.** Suppose the auctioneer in Exercise 19 sold a herd of cattle whose minimum weight was 980 pounds, median was 1140 pounds, standard deviation 84 pounds, and IQR 102 pounds. They sold for 40 cents a pound, and the auctioneer took a \$20 commission on each animal. Then, for example, a steer weighing 1100 pounds would net the owner  $0.40(1100) - 20 = \$420$ . Find the minimum, median, standard deviation, and IQR of the net sale prices.
22. **Caught speeding.** Suppose police set up radar surveillance on the Stanford street described in Exercise 18. They handed out a large number of tickets to speeders going a mean of 28 mph, with a standard deviation of 2.4 mph, a maximum of 33 mph, and an IQR of 3.2 mph. Local law prescribes fines of \$100, plus \$10 per mile per hour over the 20 mph speed limit. For example, a driver convicted of going 25 mph would be fined  $100 + 10(5) = \$150$ . Find the mean, maximum, standard deviation, and IQR of all the potential fines.

23. **Professors.** A friend tells you about a recent study dealing with the number of years of teaching experience among current college professors. He remembers the mean but can't recall whether the standard deviation was 6 months, 6 years, or 16 years. Tell him which one it must have been, and why.
24. **Rock concerts.** A popular band on tour played a series of concerts in large venues. They always drew a large crowd, averaging 21,359 fans. While the band did not announce (and probably never calculated) the standard deviation, which of these values do you think is most likely to be correct: 20, 200, 2000, or 20,000 fans? Explain your choice.
25. **Guzzlers?** Environmental Protection Agency (EPA) fuel economy estimates for automobile models tested recently predicted a mean of 24.8 mpg and a standard deviation of 6.2 mpg for highway driving. Assume that a Normal model can be applied.
- Draw the model for auto fuel economy. Clearly label it, showing what the 68–95–99.7 Rule predicts.
  - In what interval would you expect the central 68% of autos to be found?
  - About what percent of autos should get more than 31 mpg?
  - About what percent of cars should get between 31 and 37.2 mpg?
  - Describe the gas mileage of the worst 2.5% of all cars.
26. **IQ.** Some IQ tests are standardized to a Normal model, with a mean of 100 and a standard deviation of 16.
- Draw the model for these IQ scores. Clearly label it, showing what the 68–95–99.7 Rule predicts.
  - In what interval would you expect the central 95% of IQ scores to be found?
  - About what percent of people should have IQ scores above 116?
  - About what percent of people should have IQ scores between 68 and 84?
  - About what percent of people should have IQ scores above 132?
27. **Small steer.** In Exercise 17 we suggested the model  $N(1152, 84)$  for weights in pounds of yearling Angus steers. What weight would you consider to be unusually low for such an animal? Explain.
28. **High IQ.** Exercise 26 proposes modeling IQ scores with  $N(100, 16)$ . What IQ would you consider to be unusually high? Explain.
29. **Trees.** A forester measured 27 of the trees in a large woods that is up for sale. He found a mean diameter of 10.4 inches and a standard deviation of 4.7 inches. Suppose that these trees provide an accurate description of the whole forest and that a Normal model applies.
- Draw the Normal model for tree diameters.
  - What size would you expect the central 95% of all trees to be?
  - About what percent of the trees should be less than an inch in diameter?
  - About what percent of the trees should be between 5.7 and 10.4 inches in diameter?
  - About what percent of the trees should be over 15 inches in diameter?
30. **Rivets.** A company that manufactures rivets believes the shear strength (in pounds) is modeled by  $N(800, 50)$ .
- Draw and label the Normal model.
  - Would it be safe to use these rivets in a situation requiring a shear strength of 750 pounds? Explain.
  - About what percent of these rivets would you expect to fall below 900 pounds?
  - Rivets are used in a variety of applications with varying shear strength requirements. What is the maximum shear strength for which you would feel comfortable approving this company's rivets? Explain your reasoning.
31. **Trees, part II.** Later on, the forester in Exercise 29 shows you a histogram of the tree diameters he used in analyzing the woods that was for sale. Do you think he was justified in using a Normal model? Explain, citing some specific concerns.



- T 32. **Car speeds, the picture.** For the car speed data of Exercise 18, here is the histogram, boxplot, and Normal probability plot of the 100 readings. Do you think it is appropriate to apply a Normal model here? Explain.



- T 33. **Winter Olympics 2006 downhill.** Fifty-three men qualified for the men's alpine downhill race in Torino. The gold medal winner finished in 1 minute, 48.8 seconds. All competitors' times (in seconds) are found in the following list:

108.80	109.52	109.82	109.88	109.93	110.00
110.04	110.12	110.29	110.33	110.35	110.44
110.45	110.64	110.68	110.70	110.72	110.84
110.88	110.88	110.90	110.91	110.98	111.37
111.48	111.51	111.55	111.70	111.72	111.93
112.17	112.55	112.87	112.90	113.34	114.07
114.65	114.70	115.01	115.03	115.73	116.10
116.58	116.81	117.45	117.54	117.56	117.69
118.77	119.24	119.41	119.79	120.93	

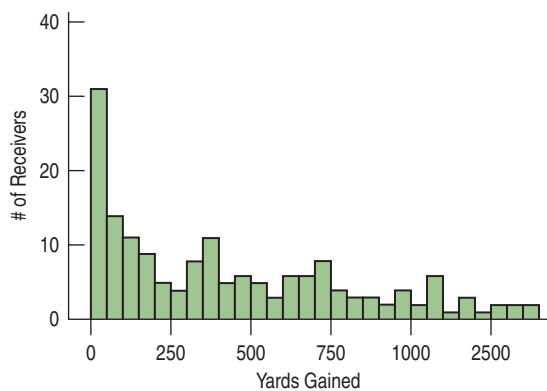
- a) The mean time was 113.02 seconds, with a standard deviation of 3.24 seconds. If the Normal model is appropriate, what percent of times will be less than 109.78 seconds?
- b) What is the actual percent of times less than 109.78 seconds?
- c) Why do you think the two percentages don't agree?
- d) Create a histogram of these times. What do you see?

- T 34. Check the model.** The mean of the 100 car speeds in Exercise 20 was 23.84 mph, with a standard deviation of 3.56 mph.
- a) Using a Normal model, what values should border the middle 95% of all car speeds?
  - b) Here are some summary statistics.

Percentile		Speed
100%	<b>Max</b>	34.060
97.5%		30.976
90.0%		28.978
75.0%	<b>Q3</b>	25.785
50.0%	<b>Median</b>	23.525
25.0%	<b>Q1</b>	21.547
10.0%		19.163
2.5%		16.638
0.0%	<b>Min</b>	16.270

From your answer in part a, how well does the model do in predicting those percentiles? Are you surprised? Explain.

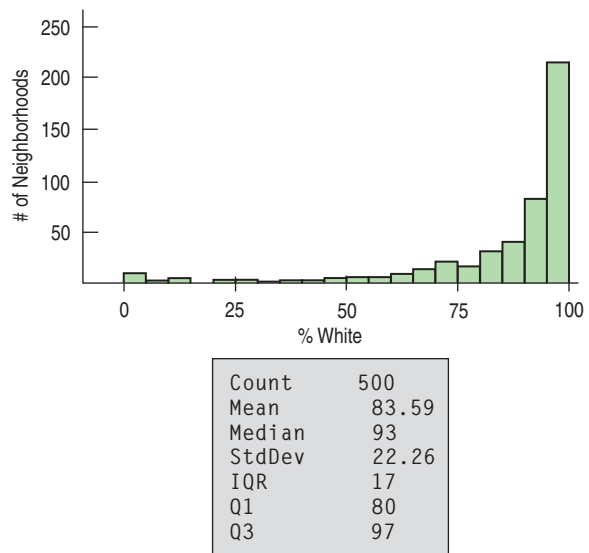
- T 35. Receivers.** NFL data from the 2006 football season reported the number of yards gained by each of the league's 167 wide receivers:



The mean is 435 yards, with a standard deviation of 384 yards.

- a) According to the Normal model, what percent of receivers would you expect to gain fewer yards than 2 standard deviations below the mean number of yards?
  - b) For these data, what does that mean?
  - c) Explain the problem in using a Normal model here.
- 36. Customer database.** A large philanthropic organization keeps records on the people who have contributed to their cause. In addition to keeping records of past giving, the organization buys demographic data on neighbor-

hoods from the U.S. Census Bureau. Eighteen of these variables concern the ethnicity of the neighborhood of the donor. Here are a histogram and summary statistics for the percentage of whites in the neighborhoods of 500 donors:



- a) Which is a better summary of the percentage of white residents in the neighborhoods, the mean or the median? Explain.
  - b) Which is a better summary of the spread, the IQR or the standard deviation? Explain.
  - c) From a Normal model, about what percentage of neighborhoods should have a percent white within one standard deviation of the mean?
  - d) What percentage of neighborhoods actually have a percent white within one standard deviation of the mean?
  - e) Explain the discrepancy between parts c and d.
- 37. Normal cattle.** Using  $N(1152, 84)$ , the Normal model for weights of Angus steers in Exercise 17, what percent of steers weigh
- a) over 1250 pounds?
  - b) under 1200 pounds?
  - c) between 1000 and 1100 pounds?
- 38. IQs revisited.** Based on the Normal model  $N(100, 16)$  describing IQ scores, what percent of people's IQs would you expect to be
- a) over 80?
  - b) under 90?
  - c) between 112 and 132?
- 39. More cattle.** Based on the model  $N(1152, 84)$  describing Angus steer weights, what are the cutoff values for
- a) the highest 10% of the weights?
  - b) the lowest 20% of the weights?
  - c) the middle 40% of the weights?
- 40. More IQs.** In the Normal model  $N(100, 16)$ , what cutoff value bounds
- a) the highest 5% of all IQs?
  - b) the lowest 30% of the IQs?
  - c) the middle 80% of the IQs?



41. **Cattle, finis.** Consider the Angus weights model  $N(1152, 84)$  one last time.
- What weight represents the 40th percentile?
  - What weight represents the 99th percentile?
  - What's the IQR of the weights of these Angus steers?
42. **IQ, finis.** Consider the IQ model  $N(100, 16)$  one last time.
- What IQ represents the 15th percentile?
  - What IQ represents the 98th percentile?
  - What's the IQR of the IQs?
43. **Cholesterol.** Assume the cholesterol levels of adult American women can be described by a Normal model with a mean of 188 mg/dL and a standard deviation of 24.
- Draw and label the Normal model.
  - What percent of adult women do you expect to have cholesterol levels over 200 mg/dL?
  - What percent of adult women do you expect to have cholesterol levels between 150 and 170 mg/dL?
  - Estimate the IQR of the cholesterol levels.
  - Above what value are the highest 15% of women's cholesterol levels?
44. **Tires.** A tire manufacturer believes that the treadlife of its snow tires can be described by a Normal model with a mean of 32,000 miles and standard deviation of 2500 miles.
- If you buy a set of these tires, would it be reasonable for you to hope they'll last 40,000 miles? Explain.
  - Approximately what fraction of these tires can be expected to last less than 30,000 miles?
  - Approximately what fraction of these tires can be expected to last between 30,000 and 35,000 miles?
  - Estimate the IQR of the treadlives.
  - In planning a marketing strategy, a local tire dealer wants to offer a refund to any customer whose tires fail to last a certain number of miles. However, the dealer does not want to take too big a risk. If the dealer is willing to give refunds to no more than 1 of every 25 customers, for what mileage can he guarantee these tires to last?
45. **Kindergarten.** Companies that design furniture for elementary school classrooms produce a variety of sizes for kids of different ages. Suppose the heights of kindergarten children can be described by a Normal model with a mean of 38.2 inches and standard deviation of 1.8 inches.
- What fraction of kindergarten kids should the company expect to be less than 3 feet tall?
  - In what height interval should the company expect to find the middle 80% of kindergarteners?
  - At least how tall are the biggest 10% of kindergarteners?
46. **Body temperatures.** Most people think that the "normal" adult body temperature is 98.6°F. That figure, based on a 19th-century study, has recently been challenged.
- In a 1992 article in the *Journal of the American Medical Association*, researchers reported that a more accurate figure may be 98.2°F. Furthermore, the standard deviation appeared to be around 0.7°F. Assume that a Normal model is appropriate.
- In what interval would you expect most people's body temperatures to be? Explain.
  - What fraction of people would be expected to have body temperatures above 98.6°F?
  - Below what body temperature are the coolest 20% of all people?
47. **Eggs.** Hens usually begin laying eggs when they are about 6 months old. Young hens tend to lay smaller eggs, often weighing less than the desired minimum weight of 54 grams.
- The average weight of the eggs produced by the young hens is 50.9 grams, and only 28% of their eggs exceed the desired minimum weight. If a Normal model is appropriate, what would the standard deviation of the egg weights be?
  - By the time these hens have reached the age of 1 year, the eggs they produce average 67.1 grams, and 98% of them are above the minimum weight. What is the standard deviation for the appropriate Normal model for these older hens?
  - Are egg sizes more consistent for the younger hens or the older ones? Explain.
48. **Tomatoes.** Agricultural scientists are working on developing an improved variety of Roma tomatoes. Marketing research indicates that customers are likely to bypass Romas that weigh less than 70 grams. The current variety of Roma plants produces fruit that averages 74 grams, but 11% of the tomatoes are too small. It is reasonable to assume that a Normal model applies.
- What is the standard deviation of the weights of Romas now being grown?
  - Scientists hope to reduce the frequency of undersized tomatoes to no more than 4%. One way to accomplish this is to raise the average size of the fruit. If the standard deviation remains the same, what target mean should they have as a goal?
  - The researchers produce a new variety with a mean weight of 75 grams, which meets the 4% goal. What is the standard deviation of the weights of these new Romas?
  - Based on their standard deviations, compare the tomatoes produced by the two varieties.



## JUST CHECKING

### Answers

1. **a)** On the first test, the mean is 88 and the SD is 4, so  $z = (90 - 88)/4 = 0.5$ . On the second test, the mean is 75 and the SD is 5, so  $z = (80 - 75)/5 = 1.0$ . The first test has the lower  $z$ -score, so it is the one that will be dropped.
  - b)** No. The second test is 1 standard deviation above the mean, farther away than the first test, so it's the better score relative to the class.
2. **a)** The mean would increase to 500.
  - b)** The standard deviation is still 100 points.
  - c)** The two boxplots would look nearly identical (the shape of the distribution would remain the same), but the later one would be shifted 50 points higher.
3. The standard deviation is now 2.54 millimeters, which is the same as 0.1 inches. Nothing has changed. The standard deviation has "increased" only because we're reporting it in millimeters now, not inches.
4. The mean is 184 centimeters, with a standard deviation of 8 centimeters. 2 meters is 200 centimeters, which is 2 standard deviations above the mean. We expect 5% of the men to be more than 2 standard deviations below or above the mean, so half of those, 2.5%, are likely to be above 2 meters.
5. **a)** We know that 68% of the time we'll be within 1 standard deviation (2 min) of 20. So 32% of the time we'll arrive in less than 18 or more than 22 minutes. Half of those times (16%) will be greater than 22 minutes, so 84% will be less than 22 minutes.
  - b)** 24 minutes is 2 standard deviations above the mean. Because of the 95% rule, we know 2.5% of the times will be more than 24 minutes.
  - c)** Traffic incidents may occasionally increase the time it takes to get to school, so the driving times may be skewed to the right, and there may be outliers.
  - d)** If so, the Normal model would not be appropriate and the percentages we predict would not be accurate.

## REVIEW OF PART I

## Exploring and Understanding Data

## Quick Review

It's time to put it all together. Real data don't come tagged with instructions for use. So let's step back and look at how the key concepts and skills we've seen work together. This brief list and the review exercises that follow should help you check your understanding of Statistics so far.

- ▶ We treat data two ways: as categorical and as quantitative.
- ▶ To describe categorical data:
  - Make a picture. Bar graphs work well for comparing counts in categories.
  - Summarize the distribution with a table of counts or relative frequencies (percents) in each category.
  - Pie charts and segmented bar charts display divisions of a whole.
  - Compare distributions with plots side by side.
  - Look for associations between variables by comparing marginal and conditional distributions.
- ▶ To describe quantitative data:
  - Make a picture. Use histograms, boxplots, stem-and-leaf displays, or dotplots. Stem-and-leaves are great when working by hand and good for small data sets. Histograms are a good way to see the distribution. Boxplots are best for comparing several distributions.
  - Describe distributions in terms of their shape, center, and spread, and note any unusual features such as gaps or outliers.
  - The shape of most distributions you'll see will likely be uniform, unimodal, or bimodal. It may be multimodal. If it is unimodal, then it may be symmetric or skewed.
  - A 5-number summary makes a good numerical description of a distribution: min, Q1, median, Q3, and max.
- If the distribution is skewed, be sure to include the median and interquartile range (IQR) when you describe its center and spread.
- A distribution that is severely skewed may benefit from re-expressing the data. If it is skewed to the high end, taking logs often works well.
- If the distribution is unimodal and symmetric, describe its center and spread with the mean and standard deviation.
- Use the standard deviation as a ruler to tell how unusual an observed value may be, or to compare or combine measurements made on different scales.
- Shifting a distribution by adding or subtracting a constant affects measures of position but not measures of spread. Rescaling by multiplying or dividing by a constant affects both.
- When a distribution is roughly unimodal and symmetric, a Normal model may be useful. For Normal models, the 68–95–99.7 Rule is a good rule of thumb.
- If the Normal model fits well (check a histogram or Normal probability plot), then Normal percentile tables or functions found in most statistics technology can provide more detailed values.

Need more help with some of this? It never hurts to reread sections of the chapters! And in the following pages we offer you more opportunities<sup>1</sup> to review these concepts and skills.

The exercises that follow use the concepts and skills you've learned in the first six chapters. To be more realistic and more useful for your review, they don't tell you which of the concepts or methods you need. But neither will the exam.

<sup>1</sup> If you doubted that we are teachers, this should convince you. Only a teacher would call additional homework exercises "opportunities."

## REVIEW EXERCISES

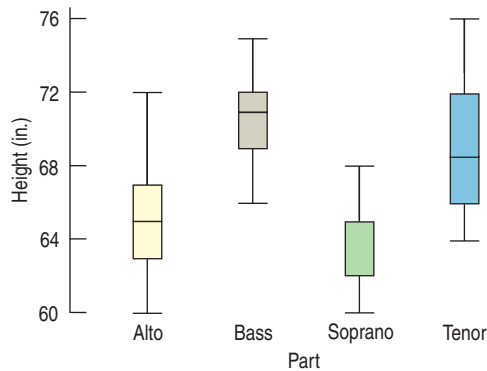
1. **Bananas.** Here are the prices (in cents per pound) of bananas reported from 15 markets surveyed by the U.S. Department of Agriculture.

51	52	45
48	53	52
50	49	52
48	43	46
45	42	50

- a) Display these data with an appropriate graph.  
 b) Report appropriate summary statistics.  
 c) Write a few sentences about this distribution.
2. **Prenatal care.** Results of a 1996 American Medical Association report about the infant mortality rate for twins carried for the full term of a normal pregnancy are shown on the next page, broken down by the level of prenatal care the mother had received.

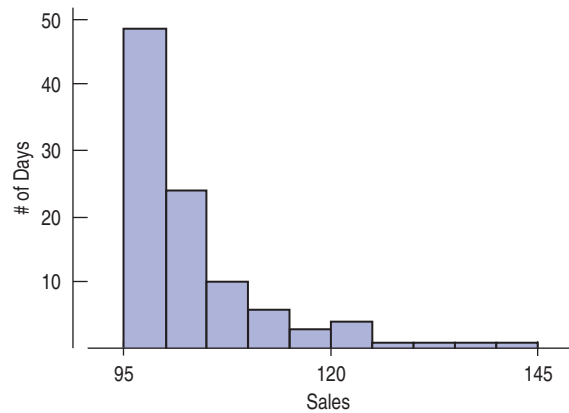
Full-Term Pregnancies, Level of Prenatal Care	Infant Mortality Rate Among Twins (deaths per thousand live births)
Intensive	5.4
Adequate	3.9
Inadequate	6.1
<b>Overall</b>	<b>5.1</b>

- Is the overall rate the average of the other three rates? Should it be? Explain.
  - Do these results indicate that adequate prenatal care is important for pregnant women? Explain.
  - Do these results suggest that a woman pregnant with twins should be wary of seeking too much medical care? Explain.
3. **Singers.** The boxplots shown display the heights (in inches) of 130 members of a choir.



- It appears that the median height for sopranos is missing, but actually the median and the upper quartile are equal. How could that happen?
  - Write a few sentences describing what you see.
4. **Dialysis.** In a study of dialysis, researchers found that “of the three patients who were currently on dialysis, 67% had developed blindness and 33% had their toes amputated.” What kind of display might be appropriate for these data? Explain.
5. **Beanstalks.** Beanstalk Clubs are social clubs for very tall people. To join, a man must be over 6’2” tall, and a woman over 5’10”. The National Health Survey suggests that heights of adults may be Normally distributed, with mean heights of 69.1” for men and 64.0” for women. The respective standard deviations are 2.8” and 2.5”.
- You are probably not surprised to learn that men are generally taller than women, but what does the greater standard deviation for men’s heights indicate?
  - Who are more likely to qualify for Beanstalk membership, men or women? Explain.

6. **Bread.** Clarksburg Bakery is trying to predict how many loaves to bake. In the last 100 days, they have sold between 95 and 140 loaves per day. Here is a histogram of the number of loaves they sold for the last 100 days.



- Describe the distribution.
- Which should be larger, the mean number of sales or the median? Explain.
- Here are the summary statistics for Clarksburg Bakery’s bread sales. Use these statistics and the histogram above to create a boxplot. You may approximate the values of any outliers.

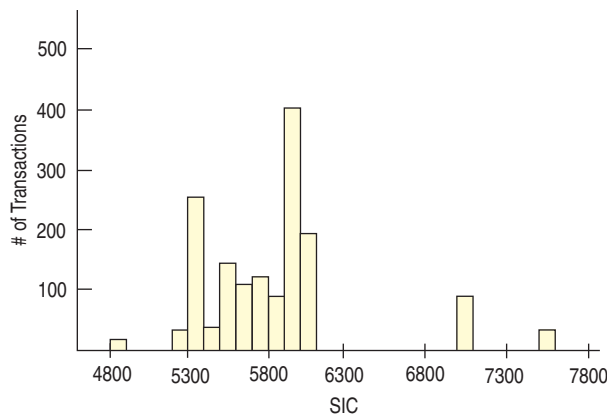
Summary of Sales	
Median	100
Min	95
Max	140
25th %tile	97
75th %tile	105.5

- For these data, the mean was 103 loaves sold per day, with a standard deviation of 9 loaves. Do these statistics suggest that Clarksburg Bakery should expect to sell between 94 and 112 loaves on about 68% of the days? Explain.
7. **State University.** Public relations staff at State U. collected data on people’s opinions of various colleges and universities in their state. They phoned 850 local residents. After identifying themselves, the callers asked the survey participants their ages, whether they had attended college, and whether they had a favorable opinion of the university. The official report to the university’s directors claimed that, in general, people had very favorable opinions about their university.
- Identify the W’s of these data.
  - Identify the variables, classify each as categorical or quantitative, and specify units if relevant.
  - Are you confident about the report’s conclusion? Explain.
8. **Acid rain.** Based on long-term investigation, researchers have suggested that the acidity (pH) of rainfall

in the Shenandoah Mountains can be described by the Normal model  $N(4.9, 0.6)$ .

- Draw and carefully label the model.
  - What percent of storms produce rainfall with pH over 6?
  - What percent of storms produce rainfall with pH under 4?
  - The lower the pH, the more acidic the rain. What is the pH level for the most acidic 20% of all storms?
  - What is the pH level for the least acidic 5% of all storms?
  - What is the IQR for the pH of rainfall?
9. **Fraud detection.** A credit card bank is investigating the incidence of fraudulent card use. The bank suspects that the type of product bought may provide clues to the fraud. To examine this situation, the bank looks at the Standard Industrial Code (SIC) of the business related to the transaction. This is a code that was used by the U.S. Census Bureau and Statistics Canada to identify the type of every registered business in North America.<sup>2</sup> For example, 1011 designates Meat and Meat Products (except Poultry), 1012 is Poultry Products, 1021 is Fish Products, 1031 is Canned and Preserved Fruits and Vegetables, and 1032 is Frozen Fruits and Vegetables.

A company intern produces the following histogram of the SIC codes for 1536 transactions:



He also reports that the mean SIC is 5823.13 with a standard deviation of 488.17.

- Comment on any problems you see with the use of the mean and standard deviation as summary statistics.
  - How well do you think the Normal model will work on these data? Explain.
10. **Streams.** As part of the course work, a class at an upstate NY college collects data on streams each year. Students record a number of biological, chemical, and physical variables, including the stream name, the substrate of the stream (*limestone, shale, or mixed*), the pH, the temperature ( $^{\circ}\text{C}$ ), and the BCI, a measure of biological diversity.

Group	Count	%
Limestone	77	44.8
Mixed	26	15.1
Shale	69	40.1

- Name each variable, indicating whether it is categorical or quantitative, and giving the units if available.
- These streams have been classified according to their substrate—the composition of soil and rock over which they flow—as summarized in the table. What kind of graph might be used to display these data?

- T 11. **Cramming.** One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Both sets of test scores for the 28 students are shown below.

Fri	Mon	Fri	Mon
42	36	50	47
44	44	34	34
45	46	38	31
48	38	43	40
44	40	39	41
43	38	46	32
41	37	37	36
35	31	40	31
43	32	41	32
48	37	48	39
43	41	37	31
45	32	36	41
47	44		

- Create a graphical display to compare the two distributions of scores.
- Write a few sentences about the scores reported on Friday and Monday.
- Create a graphical display showing the distribution of the *changes* in student scores.
- Describe the distribution of changes.

12. **Computers and Internet.** A U.S. Census Bureau report (August 2000, *Current Population Survey*) found that 51.0% of homes had a personal computer and 41.5% had access to the Internet. A newspaper concluded that 92.5% of homes had either a computer or access to the Internet. Do you agree? Explain.

13. **Let's play cards.** You pick a card from a deck (see description in Chapter 11) and record its denomination (7, say) and its suit (maybe spades).
- Is the variable *suit* categorical or quantitative?
  - Name a game you might be playing for which you would consider the variable *denomination* to be categorical. Explain.
  - Name a game you might be playing for which you would consider the variable *denomination* to be quantitative. Explain.

- T 14. **Accidents.** In 2001, Progressive Insurance asked customers who had been involved in auto accidents how far they were from home when the accident happened. The data are summarized in the table.

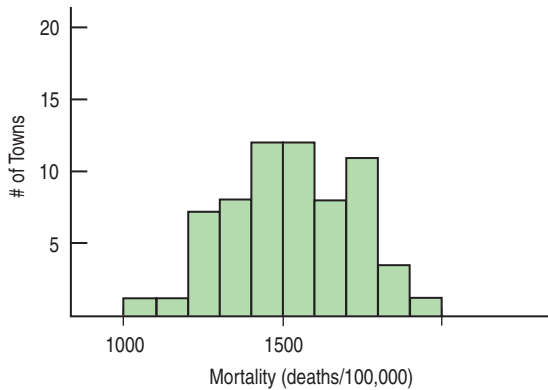
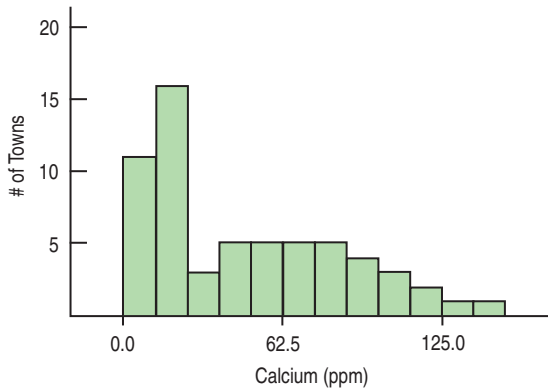
<sup>2</sup> Since 1997 the SIC has been replaced by the NAICS, a code of six letters.

Miles from Home	% of Accidents
Less than 1	23
1 to 5	29
6 to 10	17
11 to 15	8
16 to 20	6
Over 20	17

- a) Create an appropriate graph of these data.
- b) Do these data indicate that driving near home is particularly dangerous? Explain.

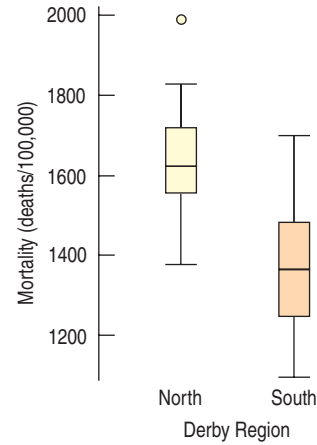
**T 15. Hard water.** In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water.

- a) What are the variables in this study? For each, indicate whether it is quantitative or categorical and what the units are.
- b) Here are histograms of calcium concentration and mortality. Describe the distributions of the two variables.



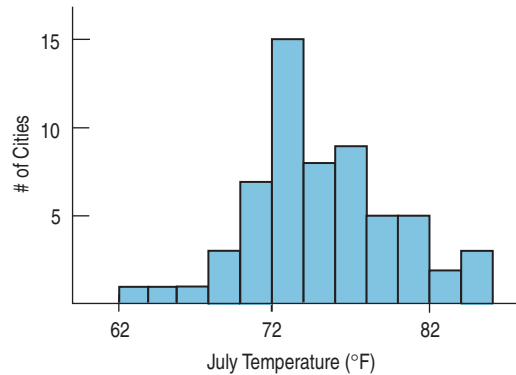
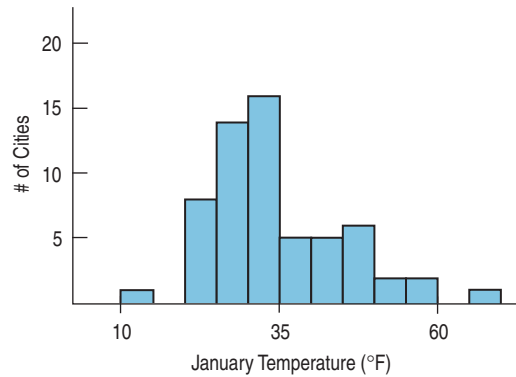
**T 16. Hard water II.** The data set from England and Wales also notes for each town whether it was south or north of Derby. Here are some summary statistics and a comparative boxplot for the two regions.

Summary of Mortality				
Group	Count	Mean	Median	StdDev
North	34	1631.59	1631	138.470
South	27	1388.85	1369	151.114

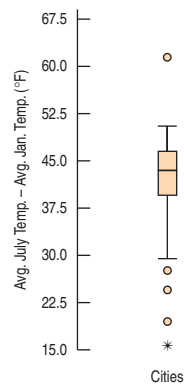


- a) What is the overall mean mortality rate for the two regions?
- b) Do you see evidence of a difference in mortality rates? Explain.

**17. Seasons.** Average daily temperatures in January and July for 60 large U.S. cities are graphed in the histograms below.

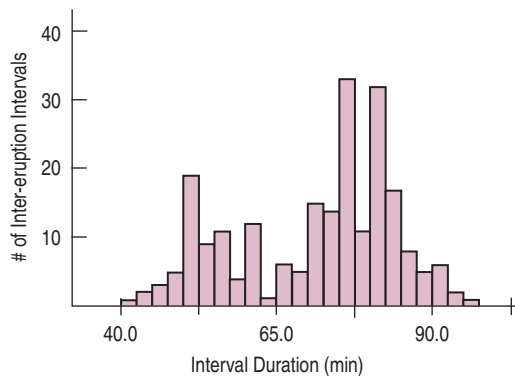


- a) What aspect of these histograms makes it difficult to compare the distributions?
- b) What differences do you see between the distributions of January and July average temperatures?



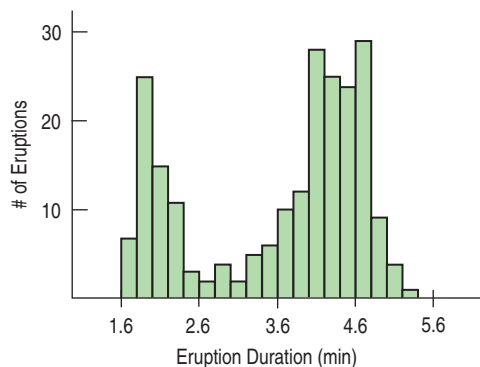
- c) Differences in temperatures (July–January) for each of the cities are displayed in the boxplot above. Write a few sentences describing what you see.

18. **Old Faithful.** It is a common belief that Yellowstone’s most famous geyser erupts once an hour at very predictable intervals. The histogram below shows the time gaps (in minutes) between 222 successive eruptions. Describe this distribution.

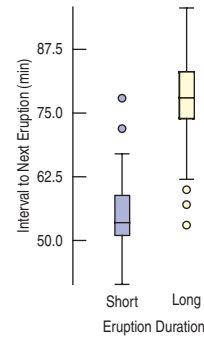


19. **Old Faithful?** Does the duration of an eruption have an effect on the length of time that elapses before the next eruption?

- a) The histogram below shows the duration (in minutes) of those 222 eruptions. Describe this distribution.



- b) Explain why it is not appropriate to find summary statistics for this distribution.
- c) Let’s classify the eruptions as “long” or “short,” depending upon whether or not they last at least 3 minutes. Describe what you see in the comparative boxplots.



20. **Teen drivers.** In its *Traffic Safety Facts 2005*, the National Highway Traffic Safety Administration reported that 6.3% of licensed drivers were between the ages of 15 and 20, yet this age group was behind the wheel in 15.9% of all fatal crashes. Use these statistics to explain the concept of independence.

- T** 21. **Liberty’s nose.** Is the Statue of Liberty’s nose too long? Her nose measures, 4’6”, but she is a large statue, after all. Her arm is 42 feet long. That means her arm is  $42/45 = 9.3$  times as long as her nose. Is that a reasonable ratio? Shown in the table are arm and nose lengths of 18 girls in a Statistics class, and the ratio of arm-to-nose length for each.

Arm (cm)	Nose (cm)	Arm/Nose Ratio
73.8	5.0	14.8
74.0	4.5	16.4
69.5	4.5	15.4
62.5	4.7	13.3
68.6	4.4	15.6
64.5	4.8	13.4
68.2	4.8	14.2
63.5	4.4	14.4
63.5	5.4	11.8
67.0	4.6	14.6
67.4	4.4	15.3
70.7	4.3	16.4
69.4	4.1	16.9
71.7	4.5	15.9
69.0	4.4	15.7
69.8	4.5	15.5
71.0	4.8	14.8
71.3	4.7	15.2

- a) Make an appropriate plot and describe the distribution of the ratios.
- b) Summarize the ratios numerically, choosing appropriate measures of center and spread.
- c) Is the ratio of 9.3 for the Statue of Liberty unrealistically low? Explain.

- T 22. Winter Olympics 2006 speed skating.** The top 25 women's 500-m speed skating times are listed in the table below:

Skater	Country	Time
Svetlana Zhurova	Russia	76.57
Wang Manli	China	76.78
Hui Ren	China	76.87
Tomomi Okazaki	Japan	76.92
Lee Sang-Hwa	South Korea	77.04
Jenny Wolf	Germany	77.25
Wang Beixing	China	77.27
Sayuri Osuga	Japan	77.39
Sayuri Yoshii	Japan	77.43
Chiara Simionato	Italy	77.68
Jennifer Rodriguez	United States	77.70
Annette Gerritsen	Netherlands	78.09
Xing Aihua	China	78.35
Sanne van der Star	Netherlands	78.59
Yukari Watanabe	Japan	78.65
Shannon Rempel	Canada	78.85
Amy Sannes	United States	78.89
Choi Seung-Yong	South Korea	79.02
Judith Hesse	Germany	79.03
Kim You-Lim	South Korea	79.25
Kerry Simpson	Canada	79.34
Krisy Myers	Canada	79.43
Elli Ochowicz	United States	79.48
Pamela Zoellner	Germany	79.56
Lee Bo-Ra	South Korea	79.73

- a) The mean finishing time was 78.21 seconds, with a standard deviation of 1.03 second. If the Normal model is appropriate, what percent of the times should be within 0.5 second of 78.21?
- b) What percent of the times actually fall within this interval?
- c) Explain the discrepancy between a and b.
- 23. Sample.** A study in South Africa focusing on the impact of health insurance identified 1590 children at birth and then sought to conduct follow-up health studies 5 years later. Only 416 of the original group participated in the 5-year follow-up study. This made researchers concerned that the follow-up group might not accurately resemble the total group in terms of health insurance. The table in the next column summarizes the two groups by race and by presence of medical insurance when the child was born. Carefully explain how this study demonstrates Simpson's paradox. (*Birth to Ten Study*, Medical Research Council, South Africa)

		Number (%) Insured	
		Follow-up	Not traced
Race	Black	36 of 404 (8.9%)	91 of 1048 (8.7%)
	White	10 of 12 (83.3%)	104 of 126 (82.5%)
	Overall	46 of 416 (11.1%)	195 of 1174 (16.6%)

- 24. Sluggers.** Roger Maris's 1961 home run record stood until Mark McGwire hit 70 in 1998. Listed below are the home run totals for each season McGwire played. Also listed are Babe Ruth's home run totals.
- McGwire:** 3\*, 49, 32, 33, 39, 22, 42, 9\*, 9\*, 39, 52, 58, 70, 65, 32\*, 29\*
- Ruth:** 54, 59, 35, 41, 46, 25, 47, 60, 54, 46, 49, 46, 41, 34, 22
- a) Find the 5-number summary for McGwire's career.
- b) Do any of his seasons appear to be outliers? Explain.
- c) McGwire played in only 18 games at the end of his first big league season, and missed major portions of some other seasons because of injuries to his back and knees. Those seasons might not be representative of his abilities. They are marked with asterisks in the list above. Omit these values and make parallel boxplots comparing McGwire's career to Babe Ruth's.
- d) Write a few sentences comparing the two sluggers.
- e) Create a side-by-side stem-and-leaf display comparing the careers of the two players.
- f) What aspects of the distributions are apparent in the stem-and-leaf displays that did not clearly show in the boxplots?
- 25. Be quick!** Avoiding an accident when driving can depend on reaction time. That time, measured from the moment the driver first sees the danger until he or she steps on the brake pedal, is thought to follow a Normal model with a mean of 1.5 seconds and a standard deviation of 0.18 seconds.
- a) Use the 68–95–99.7 Rule to draw the Normal model.
- b) Write a few sentences describing driver reaction times.
- c) What percent of drivers have a reaction time less than 1.25 seconds?
- d) What percent of drivers have reaction times between 1.6 and 1.8 seconds?
- e) What is the interquartile range of reaction times?
- f) Describe the reaction times of the slowest 1/3 of all drivers.
- 26. Music and memory.** Is it a good idea to listen to music when studying for a big test? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, Mozart, or no music



while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are the 5-number summaries for each group:

	<i>n</i>	Min	Q1	Median	Q3	Max
<b>Rap</b>	29	5	8	10	12	25
<b>Mozart</b>	20	4	7	10	12	27
<b>None</b>	13	8	9.5	13	17	24

- Describe the *W*'s for these data: *Who, What, Where, Why, When, How*.
- Name the variables and classify each as categorical or quantitative.
- Create parallel boxplots as best you can from these summary statistics to display these results.
- Write a few sentences comparing the performances of the three groups.

- T 27. Mail.** Here are the number of pieces of mail received at a school office for 36 days.

123	70	90	151	115	97
80	78	72	100	128	130
52	103	138	66	135	76
112	92	93	143	100	88
118	118	106	110	75	60
95	131	59	115	105	85

- Plot these data.
- Find appropriate summary statistics.
- Write a brief description of the school's mail deliveries.
- What percent of the days actually lie within one standard deviation of the mean? Comment.

- T 28. Birth order.** Is your birth order related to your choice of major? A Statistics professor at a large university polled his students to find out what their majors were and what position they held in the family birth order. The results are summarized in the table.

- What percent of these students are oldest or only children?
- What percent of Humanities majors are oldest children?
- What percent of oldest children are Humanities students?
- What percent of the students are oldest children majoring in the Humanities?

		Birth Order*				Total
		1	2	3	4+	
Major	Math/Science	34	14	6	3	57
	Agriculture	52	27	5	9	93
	Humanities	15	17	8	3	43
	Other	12	11	1	6	30
	<b>Total</b>	<b>113</b>	<b>69</b>	<b>20</b>	<b>21</b>	<b>223</b>

\* 1 = oldest or only child

- 29. Herbal medicine.** Researchers for the Herbal Medicine Council collected information on people's experiences with a new herbal remedy for colds. They went to a store that sold natural health products. There they asked 100 customers whether they had taken the cold remedy and, if so, to rate its effectiveness (on a scale from 1 to 10) in curing their symptoms. The Council concluded that this product was highly effective in treating the common cold.
- Identify the *W*'s of these data.
  - Identify the variables, classify each as categorical or quantitative, and specify units if relevant.
  - Are you confident about the Council's conclusion? Explain.

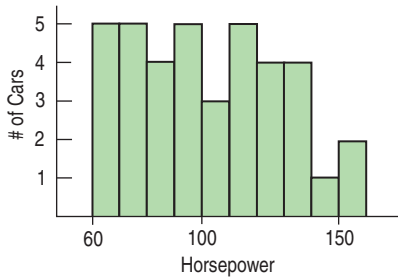
- T 30. Birth order revisited.** Consider again the data on birth order and college majors in Exercise 28.
- What is the marginal distribution of majors?
  - What is the conditional distribution of majors for the oldest children?
  - What is the conditional distribution of majors for the children born second?
  - Do you think that college major appears to be independent of birth order? Explain.

- 31. Engines.** One measure of the size of an automobile engine is its "displacement," the total volume (in liters or cubic inches) of its cylinders. Summary statistics for several models of new cars are shown. These displacements were measured in cubic inches.

Summary of Displacement	
Count	38
Mean	177.29
Median	148.5
StdDev	88.88
Range	275
25th %tile	105
75th %tile	231

- How many cars were measured?
  - Why might the mean be so much larger than the median?
  - Describe the center and spread of this distribution with appropriate statistics.
  - Your neighbor is bragging about the 227-cubic-inch engine he bought in his new car. Is that engine unusually large? Explain.
  - Are there any engines in this data set that you would consider to be outliers? Explain.
  - Is it reasonable to expect that about 68% of car engines measure between 88 and 266 cubic inches? (That's  $177.289 \pm 88.8767$ .) Explain.
  - We can convert all the data from cubic inches to cubic centimeters (cc) by multiplying by 16.4. For example, a 200-cubic-inch engine has a displacement of 3280 cc. How would such a conversion affect each of the summary statistics?
- 32. Engines, again.** Horsepower is another measure commonly used to describe auto engines. Here are the summary statistics and histogram displaying horsepowers of the same group of 38 cars discussed in Exercise 31.

Summary of Horsepower	
Count	38
Mean	101.7
Median	100
StdDev	26.4
Range	90
25th %tile	78
75th %tile	125



- Describe the shape, center, and spread of this distribution.
  - What is the interquartile range?
  - Are any of these engines outliers in terms of horsepower? Explain.
  - Do you think the 68–95–99.7 Rule applies to the horsepower of auto engines? Explain.
  - From the histogram, make a rough estimate of the percentage of these engines whose horsepower is within one standard deviation of the mean.
  - A fuel additive boasts in its advertising that it can “add 10 horsepower to any car.” Assuming that is true, what would happen to each of these summary statistics if this additive were used in all the cars?
33. **Age and party 2007.** The Pew Research Center conducts surveys regularly asking respondents which political party they identify with. Among their results is the following table relating preferred political party and age. (<http://people-press.org/reports/>)

	Party			Total
	Republican	Democrat	Others	
Age				
18–29	2636	2738	4765	10139
30–49	6871	6442	8160	21473
50–64	3896	4286	4806	12988
65+	3131	3718	2934	9784
Total	16535	17183	20666	54384

- What percent of people surveyed were Republicans?
- Do you think this might be a reasonable estimate of the percentage of all voters who are Republicans? Explain.
- What percent of people surveyed were under 30 or over 65?
- What percent of people were classified as “Other” and under the age of 30?

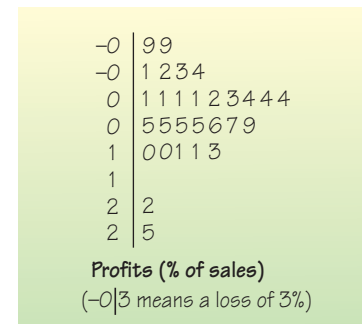
- What percent of the people classified as “Other” were under 30?
  - What percent of people under 30 were classified as “Other”?
34. **Pay.** According to the 2006 *National Occupational Employment and Wage Estimates for Management Occupations*, the mean hourly wage for Chief Executives was \$69.52 and the median hourly wage was “over \$70.00.” By contrast, for General and Operations Managers, the mean hourly wage was \$47.73 and the median was \$40.97. Are these wage distributions likely to be symmetric, skewed left, or skewed right? Explain.
35. **Age and party II.** Consider again the Pew Research Center results on age and political party in Exercise 33.
- What is the marginal distribution of party affiliation?
  - Create segmented bar graphs displaying the conditional distribution of party affiliation for each age group.
  - Summarize these poll results in a few sentences that might appear in a newspaper article about party affiliation in the United States.
  - Do you think party affiliation is independent of the voter’s age? Explain.
- T 36. **Bike safety 2003.** The Bicycle Helmet Safety Institute website includes a report on the number of bicycle fatalities per year in the United States. The table below shows the counts for the years 1994–2003.

Year	Bicycle fatalities
1994	796
1995	828
1996	761
1997	811
1998	757
1999	750
2000	689
2001	729
2002	663
2003	619

- What are the W’s for these data?
  - Display the data in a stem-and-leaf display.
  - Display the data in a timeplot.
  - What is apparent in the stem-and-leaf display that is hard to see in the timeplot?
  - What is apparent in the timeplot that is hard to see in the stem-and-leaf display?
  - Write a few sentences about bicycle fatalities in the United States.
37. **Some assembly required.** A company that markets build-it-yourself furniture sells a computer desk that is advertised with the claim “less than an hour to assemble.” However, through postpurchase surveys the company has learned that only 25% of its customers succeeded in building the desk in under an hour. The mean time was 1.29 hours. The company assumes that consumer assembly time follows a Normal model.

- Find the standard deviation of the assembly time model.
- One way the company could solve this problem would be to change the advertising claim. What assembly time should the company quote in order that 60% of customers succeed in finishing the desk by then?
- Wishing to maintain the “less than an hour” claim, the company hopes that revising the instructions and labeling the parts more clearly can improve the 1-hour success rate to 60%. If the standard deviation stays the same, what new lower mean time does the company need to achieve?
- Months later, another postpurchase survey shows that new instructions and part labeling did lower the mean assembly time, but only to 55 minutes. Nonetheless, the company did achieve the 60%-in-an-hour goal, too. How was that possible?

**T** 38. **Profits.** Here is a stem-and-leaf display showing profits as a percent of sales for 29 of the *Forbes* 500 largest U.S. corporations. The stems are split; each stem represents a span of 5%, from a loss of 9% to a profit of 25%.



- Find the 5-number summary.
- Draw a boxplot for these data.
- Find the mean and standard deviation.
- Describe the distribution of profits for these corporations.



PART

# Exploring Relationships Between Variables

## Chapter 7

Scatterplots, Association, and Correlation

## Chapter 8

Linear Regression

## Chapter 9

Regression Wisdom

## Chapter 10

Re-expressing Data: Get It Straight!

# Scatterplots, Association, and Correlation



**WHO** Years 1970–2005

**WHAT** Mean error in the position of Atlantic hurricanes as predicted 72 hours ahead by the NHC

**UNITS** nautical miles

**WHEN** 1970–2005

**WHERE** Atlantic and Gulf of Mexico

**WHY** The NHC wants to improve prediction models

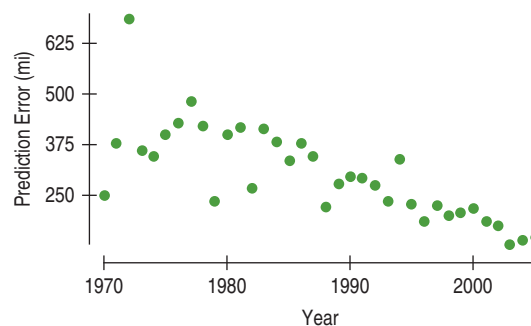
### Look, Ma, no origin!

Scatterplots usually don't—and shouldn't—show the origin, because often neither variable has values near 0. The display should focus on the part of the coordinate plane that actually contains the data. In our example about hurricanes, none of the prediction errors or years were anywhere near 0, so the computer drew the scatterplot with axes that don't quite meet.

**H**urricane Katrina killed 1,836 people<sup>1</sup> and caused well over 100 billion dollars in damage—the most ever recorded. Much of the damage caused by Katrina was due to its almost perfectly deadly aim at New Orleans.

Where will a hurricane go? People want to know if a hurricane is coming their way, and the National Hurricane Center (NHC) of the National Oceanic and Atmospheric Administration (NOAA) tries to predict the path a hurricane will take. But hurricanes tend to wander around aimlessly and are pushed by fronts and other weather phenomena in their area, so they are notoriously difficult to predict. Even relatively small changes in a hurricane's track can make big differences in the damage it causes.

To improve hurricane prediction, NOAA<sup>2</sup> relies on sophisticated computer models, and has been working for decades to improve them. How well are they doing? Have predictions improved in recent years? Has the improvement been consistent? Here's a timeplot of the mean error, in nautical miles, of the NHC's 72-hour predictions of Atlantic hurricanes since 1970:



**FIGURE 7.1**

A scatterplot of the average error in nautical miles of the predicted position of Atlantic hurricanes for predictions made by the National Hurricane Center of NOAA, plotted against the Year in which the predictions were made.

<sup>1</sup> In addition, 705 are still listed as missing.

<sup>2</sup> [www.nhc.noaa.gov](http://www.nhc.noaa.gov)

**AS** **Activity: Heights of Husbands and Wives.** Husbands are usually taller than their wives. Or are they?

Clearly, predictions have improved. The plot shows a fairly steady decline in the average error, from almost 500 nautical miles in the late 1970s to about 150 nautical miles in 2005. We can also see a few years when predictions were unusually good and that 1972 was a really bad year for predicting hurricane tracks.

This timeplot is an example of a more general kind of display called a **scatterplot**. Scatterplots may be the most common displays for data. By just looking at them, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others. As the great philosopher Yogi Berra<sup>3</sup> once said, “You can observe a lot by watching.”<sup>4</sup> Scatterplots are the best way to start observing the relationship between two *quantitative* variables.

Relationships between variables are often at the heart of what we’d like to learn from data:

- ▶ Are grades actually higher now than they used to be?
- ▶ Do people tend to reach puberty at a younger age than in previous generations?
- ▶ Does applying magnets to parts of the body relieve pain? If so, are stronger magnets more effective?
- ▶ Do students learn better with more use of computer technology?

Questions such as these relate two quantitative variables and ask whether there is an **association** between them. Scatterplots are the ideal way to *picture* such associations.

## Looking at Scatterplots



**AS** **Activity: Making and Understanding Scatterplots.** See the best way to make scatterplots—using a computer.

Look for **Direction**: What’s my sign—positive, negative, or neither?

Look for **Form**: straight, curved, something exotic, or no pattern?


How would you describe the association of hurricane *Prediction Error* and *Year*? Everyone looks at scatterplots. But, if asked, many people would find it hard to say what to look for in a scatterplot. What do *you* see? Try to describe the scatterplot of *Prediction Error* against *Year*.


You might say that the **direction** of the association is important. Over time, the NHC’s prediction errors have decreased. A pattern like this that runs from the

upper left to the lower right  is said to be **negative**. A pattern running the other way  is called **positive**.

The second thing to look for in a scatterplot is its **form**. If there is a straight line relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form. For example, the scatterplot of *Prediction Error* vs. *Year* has such an underlying **linear** form, although some points stray away from it.

Scatterplots can reveal many kinds of patterns. Often they will not be straight, but straight line patterns are both the most common and the most useful for statistics.

If the relationship isn’t straight, but curves gently, while still increasing or decreasing steadily, , we can often find ways to make it more nearly

straight. But if it curves sharply—up and then down, for example —there is much less we can say about it with the methods of this book.

<sup>3</sup> Hall of Fame catcher and manager of the New York Mets and Yankees.

<sup>4</sup> But then he also said “I really didn’t say everything I said.” So we can’t really be sure.

Look for **Strength**: how much scatter?

The third feature to look for in a scatterplot is how strong the relationship is.

At one extreme, do the points appear tightly clustered in a single stream (whether straight, curved, or bending all over the place)? Or, at the other extreme, does the swarm of points seem to form a vague cloud through which we can



barely discern any trend or pattern?

The *Prediction error vs. Year*

plot shows moderate scatter around a generally straight form. This indicates that the linear trend of improving prediction is pretty consistent and moderately strong.

Look for **Unusual Features**: Are there outliers or subgroups?

Finally, always look for the unexpected. Often the most interesting thing to see in a scatterplot is something you never thought to look for. One example of such a surprise is an **outlier** standing away from the overall pattern of the scatterplot. Such a point is almost always interesting and always deserves special attention. In the scatterplot of prediction errors, the year 1972 stands out as a year with very high prediction errors. An Internet search shows that it was a relatively quiet hurricane season. However, it included the very unusual—and deadly—Hurricane Agnes, which combined with another low-pressure center to ravage the northeastern United States, killing 122 and causing 1.3 billion 1972 dollars in damage. Possibly, Agnes was also unusually difficult to predict.

You should also look for clusters or subgroups that stand away from the rest of the plot or that show a trend in a different direction. Deviating groups should raise questions about why they are different. They may be a clue that you should split the data into subgroups instead of looking at them all together.

## FOR EXAMPLE

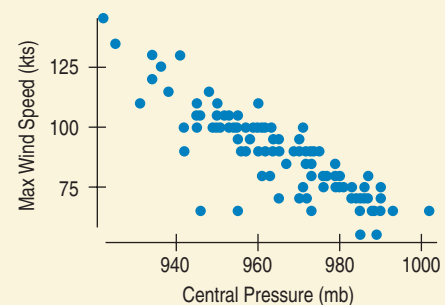
### Describing the scatterplot of hurricane winds and pressure

Hurricanes develop low pressure at their centers. This pulls in moist air, pumps up their rotation, and generates high winds. Standard sea-level pressure is around 1013 millibars (mb), or 29.9 inches of mercury. Hurricane Katrina had a central pressure of 920 mb and sustained winds of 110 knots.

Here's a scatterplot of *Maximum Wind Speed (kts) vs. Central Pressure (mb)* for 163 hurricanes that have hit the United States since 1851.

**Question:** Describe what this plot shows.

The scatterplot shows a negative direction; in general, lower central pressure is found in hurricanes that have higher maximum wind speeds. This association is linear and moderately strong.



## Roles for Variables

Which variable should go on the  $x$ -axis and which on the  $y$ -axis? What we want to know about the relationship can tell us how to make the plot. We often have questions such as:

- ▶ Do baseball teams that score more runs sell more tickets to their games?
- ▶ Do older houses sell for less than newer ones of comparable size and quality?

- ▶ Do students who score higher on their SAT tests have higher grade point averages in college?
- ▶ Can we estimate a person's percent body fat more simply by just measuring waist or wrist size?

### NOTATION ALERT

So  $x$  and  $y$  are reserved letters as well, but not just for labeling the axes of a scatterplot. In Statistics, the assignment of variables to the  $x$ - and  $y$ -axes (and the choice of notation for them in formulas) often conveys information about their roles as predictor or response variable.

#### AS Self-Test: Scatterplot

**Check.** Can you identify a scatterplot's direction, form, and strength?

In these examples, the two variables play different roles. We'll call the variable of interest the **response variable** and the other the **explanatory** or **predictor variable**.<sup>5</sup> We'll continue our practice of naming the variable of interest  $y$ . Naturally we'll plot it on the  $y$ -axis and place the explanatory variable on the  $x$ -axis. Sometimes, we'll call them the  **$x$ - and  $y$ -variables**. When you make a scatterplot, you can assume that those who view it will think this way, so choose which variables to assign to which axes carefully.

The roles that we choose for variables are more about how we *think* about them than about the variables themselves. Just placing a variable on the  $x$ -axis doesn't necessarily mean that it explains or predicts *anything*. And the variable on the  $y$ -axis may not respond to it in any way. We plotted prediction error on the  $y$ -axis against year on the  $x$ -axis because the National Hurricane Center is interested in how their predictions have changed over time. Could we have plotted them the other way? In this case, it's hard to imagine reversing the roles—knowing the prediction error and wanting to guess in what year it happened. But for some scatterplots, it can make sense to use either choice, so you have to think about how the choice of role helps to answer the question you have.

### TI Tips

### Creating a scatterplot

Let's use your calculator to make a scatterplot. First you need some data. It's okay to just enter the data in any two lists, but let's get fancy. When you are handling lots of data and several variables (as you will be soon), remembering what you stored in **L1**, **L2**, and so on can become confusing. You can—and should—give your variables meaningful names. To see how, let's store some data that you will use several times in this chapter and the next. They show the change in tuition costs at Arizona State University during the 1990s.

#### Naming the Lists

- Go into **STAT Edit**, place the cursor on one of the list names (**L1**, say), and use the arrow key to move to the right across all the lists until you encounter a blank column.
- Type **YR** to name this first variable, then hit **ENTER**.
- Often when we work with years it makes sense to use values like "90" (or even "0") rather than big numbers like "1990." For these data enter the years 1990 through 2000 as 0, 1, 2, . . . , 10.
- Now go to the next blank column, name this variable **TUIT**, and enter these values: 6546, 6996, 6996, 7350, 7500, 7978, 8377, 8710, 9110, 9411, 9800.

YR	TUIT	
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

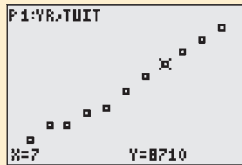
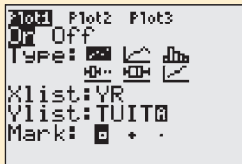
Name=TUIT

YR	TUIT	
0	6546	
1	6996	
2	6996	
3	7350	
4	7500	
5	7978	
6	8377	
7		
8		
9		
10		

Name=

<sup>5</sup> The  $x$ - and  $y$ -variables have sometimes been referred to as the *independent* and *dependent* variables, respectively. The idea was that the  $y$ -variable depended on the  $x$ -variable and the  $x$ -variable acted independently to make  $y$  respond. These names, however, conflict with other uses of the same terms in Statistics.





### Making the Scatterplot

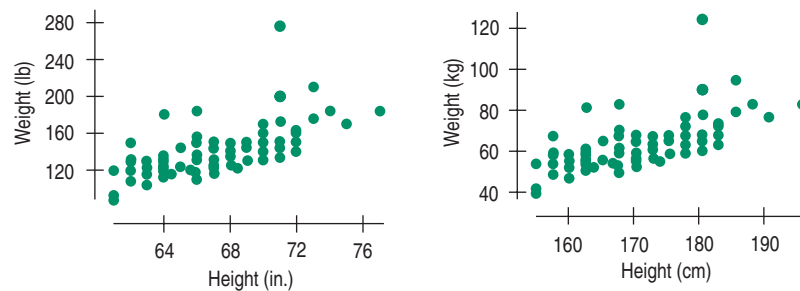
- Set up the **STATPLOT** by choosing the scatterplot icon (the first option).
- Identify which lists you want as **Xlist** and **Ylist**. If the data are in **L1** and **L2**, that's easy to do—but your data are stored in lists with special names. To specify your **Xlist**, go to **2nd LIST NAMES**, scroll down the list of variables until you find **YR**, then hit **ENTER**.
- Use **LIST NAMES** again to specify **Ylist:TUIT**.
- Pick a symbol for displaying the points.
- Now **ZoomStat** to see your scatterplot. (Didn't work? **ERR: DIM MISMATCH** means you don't have the same number of  $x$ 's and  $y$ 's. Go to **STAT Edit** and look carefully at your two datalists. You can easily fix the problem once you find it.)
- Notice that if you **TRACE** the scatterplot the calculator will tell you the  $x$ - and  $y$ -value at each point.

What can you Tell about the trend in tuition costs at ASU? (Remember: direction, form, and strength!)

## Correlation

<b>WHO</b>	Students
<b>WHAT</b>	Height (inches), weight (pounds)
<b>WHERE</b>	Ithaca, NY
<b>WHY</b>	Data for class
<b>HOW</b>	Survey

Data collected from students in Statistics classes included their *Height* (in inches) and *Weight* (in pounds). It's no great surprise to discover that there is a positive association between the two. As you might suspect, taller students tend to weigh more. (If we had reversed the roles and chosen height as the explanatory variable, we might say that heavier students tend to be taller.)<sup>6</sup> And the form of the scatterplot is fairly straight as well, although there seems to be a high outlier, as the plot shows.



**FIGURE 7.2** *Weight vs. Height of Statistics students.*

*Plotting Weight vs. Height in different units doesn't change the shape of the pattern.*

### Activity: Correlation.

Here's a good example of how correlation works to summarize the strength of a linear relationship and disregard scaling.

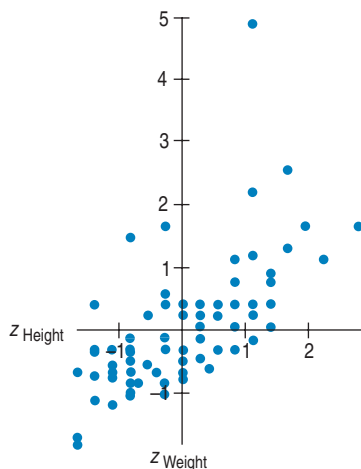
The pattern in the scatterplots looks straight and is clearly a positive association, but how strong is it? If you had to put a number (say, between 0 and 1) on the strength, what would it be? Whatever measure you use shouldn't depend on the choice of units for the variables. After all, if we measure heights and weights in centimeters and kilograms instead, it doesn't change the direction, form, or strength, so it shouldn't change the number.

<sup>6</sup> The son of one of the authors, when told (as he often was) that he was tall for his age, used to point out that, actually, he was young for his height.

Since the units shouldn't matter to our measure of strength, we can remove them by standardizing each variable. Now, for each point, instead of the values  $(x, y)$  we'll have the standardized coordinates  $(z_x, z_y)$ . Remember that to standardize values, we subtract the mean of each variable and then divide by its standard deviation:

$$(z_x, z_y) = \left( \frac{x - \bar{x}}{s_x}, \frac{y - \bar{y}}{s_y} \right).$$

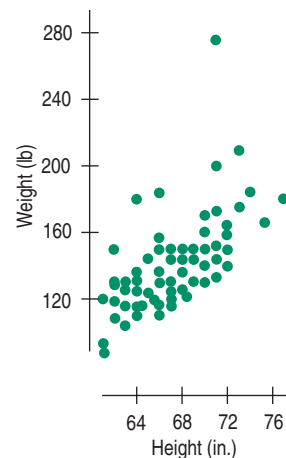
Because standardizing makes the means of both variables 0, the center of the new scatterplot is at the origin. The scales on both axes are now standard deviation units.



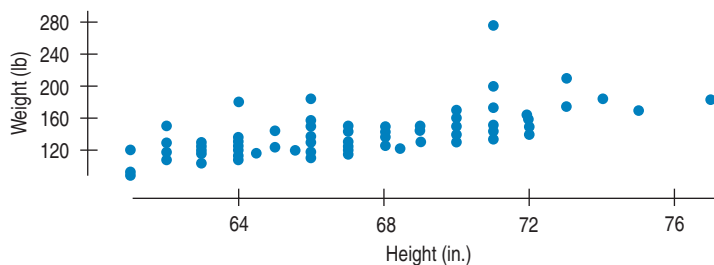
**FIGURE 7.3**

*A scatterplot of standardized heights and weights.*

Standardizing shouldn't affect the appearance of the plot. Does the plot of z-scores (Figure 7.3) look like the previous plots? Well, no. The underlying linear pattern seems steeper in the standardized plot. That's because the scales of the axes are now the same, so the length of one standard deviation is the same vertically and horizontally. When we worked in the original units, we were free to make the plot as tall and thin



or as squat and wide



as we wanted to, but that can change the impression the plot gives. By contrast,

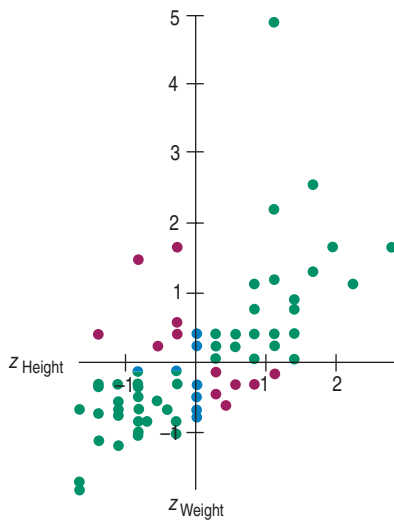


FIGURE 7.4

In this scatterplot of z-scores, points are colored according to how they affect the association: green for positive, red for negative, and blue for neutral.

**AS** **Activity: Correlation and Relationship Strength.** What does a correlation of 0.8 look like? How about 0.3?

### NOTATION ALERT

The letter  $r$  is always used for correlation, so you can't use it for anything else in Statistics. Whenever you see an  $r$ , it's safe to assume it's a correlation.

**AS** **Simulation: Correlation and Linearity.** How much does straightness matter?

equal scaling gives a neutral way of drawing the scatterplot and a fairer impression of the strength of the association.<sup>7</sup>

Which points in the scatterplot of the z-scores give the impression of a positive association? In a positive association,  $y$  tends to increase as  $x$  increases. So, the points in the upper right and lower left (colored green) strengthen that impression. For these points,  $z_x$  and  $z_y$  have the same sign, so the product  $z_x z_y$  is positive. Points far from the origin (which make the association look more positive) have bigger products.

The red points in the upper left and lower right quadrants tend to weaken the positive association (or support a negative association). For these points,  $z_x$  and  $z_y$  have opposite signs. So the product  $z_x z_y$  for these points is negative. Points far from the origin (which make the association look more negative) have a negative product even larger in magnitude.

Points with z-scores of zero on either variable don't vote either way, because  $z_x z_y = 0$ . They're colored blue.

To turn these products into a measure of the strength of the association, just add up the  $z_x z_y$  products for every point in the scatterplot:

$$\sum z_x z_y.$$

This summarizes the direction *and* strength of the association for all the points. If most of the points are in the green quadrants, the sum will tend to be positive. If most are in the red quadrants, it will tend to be negative.

But the *size* of this sum gets bigger the more data we have. To adjust for this, the natural (for statisticians anyway) thing to do is to divide the sum by  $n - 1$ .<sup>8</sup> The ratio is the famous **correlation coefficient**:

$$r = \frac{\sum z_x z_y}{n - 1}.$$

For the students' heights and weights, the correlation is 0.644. There are a number of alternative formulas for the correlation coefficient, but this form using z-scores is best for understanding what correlation means.

## Correlation Conditions

**Correlation** measures the strength of the *linear* association between two *quantitative* variables. Before you use correlation, you must check several *conditions*:

- ▶ **Quantitative Variables Condition:** Are both variables quantitative? Correlation applies only to quantitative variables. Don't apply correlation to categorical data masquerading as quantitative. Check that you know the variables' units and what they measure.
- ▶ **Straight Enough Condition:** Is the form of the scatterplot straight enough that a linear relationship makes sense? Sure, you can *calculate* a correlation coefficient for any pair of variables. But correlation measures the strength only

<sup>7</sup> When we draw a scatterplot, what often looks best is to make the length of the  $x$ -axis slightly larger than the length of the  $y$ -axis. This is an aesthetic choice, probably related to the Golden Ratio of the Greeks.

<sup>8</sup> Yes, the same  $n - 1$  as in the standard deviation calculation. And we offer the same promise to explain it later.

AS

**Case Study: Mortality and Education.** Is the mortality rate lower in cities with higher education levels?

of the *linear* association, and will be misleading if the relationship is not linear. What is “straight enough”? How non-straight would the scatterplot have to be to fail the condition? This is a judgment call that you just have to think about. Do you think that the underlying relationship is curved? If so, then summarizing its strength with a correlation would be misleading.

- ▶ **Outlier Condition:** Outliers can distort the correlation dramatically. An outlier can make an otherwise weak correlation look big or hide a strong correlation. It can even give an otherwise positive association a negative correlation coefficient (and vice versa). When you see an outlier, it’s often a good idea to report the correlation with and without that point.

Each of these conditions is easy to check with a scatterplot. Many correlations are reported without supporting data or plots. Nevertheless, you should still think about the conditions. And you should be cautious in interpreting (or accepting others’ interpretations of) the correlation when you can’t check the conditions for yourself.

## FOR EXAMPLE

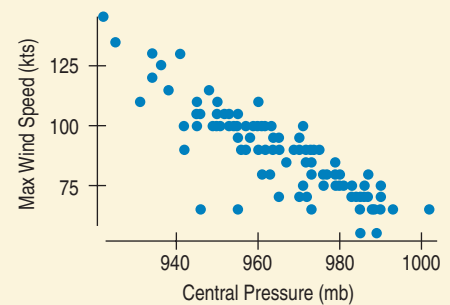
### Correlating wind speed and pressure

**Recap:** We looked at the scatterplot displaying hurricane wind speeds and central pressures.

The correlation coefficient for these wind speeds and pressures is  $r = -0.879$ .

**Question:** Check the conditions for using correlation. If you feel they are satisfied, interpret this correlation.

- ▶ **Quantitative Variables Condition:** Both wind speed and central pressure are quantitative variables, measured (respectively) in knots and millibars.
- ▶ **Straight Enough Condition:** The pattern in the scatterplot is quite straight.
- ▶ **Outlier Condition:** A few hurricanes seem to straggle away from the main pattern, but they don’t appear to be extreme enough to be called outliers. It may be worthwhile to check on them, however.



The conditions for using correlation are satisfied. The correlation coefficient of  $r = -0.879$  indicates quite a strong negative linear association between the wind speeds of hurricanes and their central pressures.



## JUST CHECKING

Your Statistics teacher tells you that the correlation between the scores (points out of 50) on Exam 1 and Exam 2 was 0.75.

1. Before answering any questions about the correlation, what would you like to see? Why?
2. If she adds 10 points to each Exam 1 score, how will this change the correlation?
3. If she standardizes scores on each exam, how will this affect the correlation?
4. In general, if someone did poorly on Exam 1, are they likely to have done poorly or well on Exam 2? Explain.
5. If someone did poorly on Exam 1, can you be sure that they did poorly on Exam 2 as well? Explain.

## STEP-BY-STEP EXAMPLE

## Looking at Association

When your blood pressure is measured, it is reported as two values: systolic blood pressure and diastolic blood pressure.

**Questions:** How are these variables related to each other? Do they tend to be both high or both low? How strongly associated are they?

THINK

**Plan** State what you are trying to investigate.

**Variables** Identify the two quantitative variables whose relationship we wish to examine. Report the *W*'s, and be sure both variables are recorded for the same individuals.

**Plot** Make the scatterplot. Use a computer program or graphing calculator if you can.

Check the conditions.

REALITY CHECK

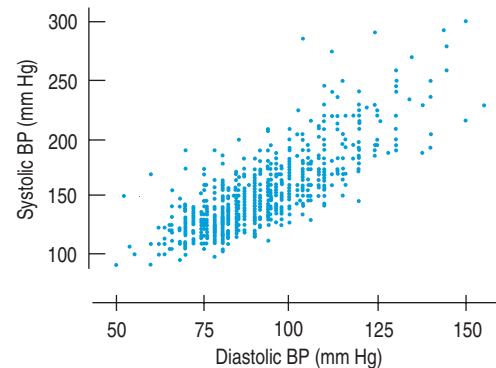
Looks like a strong positive linear association. We shouldn't be surprised if the correlation coefficient is positive and fairly large.

SHOW

**Mechanics** We usually calculate correlations with technology. Here we have 1406 cases, so we'd never try it by hand.

I'll examine the relationship between two measures of blood pressure.

The variables are systolic and diastolic blood pressure (SBP and DBP), recorded in millimeters of mercury (mm Hg) for each of 1406 participants in the Framingham Heart Study, a famous health study in Framingham, MA.<sup>9</sup>



- ✓ **Quantitative Variables Condition:** Both SBP and DBP are quantitative and measured in mm Hg.
- ✓ **Straight Enough Condition:** The scatterplot looks straight.
- ✓ **Outlier Condition:** There are a few straggling points, but none far enough from the body of the data to be called outliers.

I have two quantitative variables that satisfy the conditions, so correlation is a suitable measure of association.

The correlation coefficient is  $r = 0.792$ .

<sup>9</sup> [www.nhlbi.nih.gov/about/framingham](http://www.nhlbi.nih.gov/about/framingham)



**Conclusion** Describe the direction, form, and strength you see in the plot, along with any unusual points or features. Be sure to state your interpretations in the proper context.

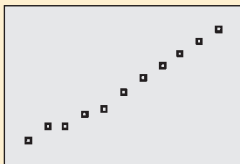
The scatterplot shows a positive direction, with higher *SBP* going with higher *DBP*. The plot is generally straight, with a moderate amount of scatter. The correlation of 0.792 is consistent with what I saw in the scatterplot. A few cases stand out with unusually high *SBP* compared with their *DBP*. It seems far less common for the *DBP* to be high by itself.

## TI Tips

## Finding the correlation

```
CATALOG
DependAuto
det(
DiagnosticOff
DiagnosticOn
dim(
Disp
DispGraph
```

```
DiagnosticOn Done
█
```



```
EDIT 0:CALC TESTS
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg
8:LinReg(a+bx)
9:LnReg
0:ExpReg
```

```
LinReg(a+bx) L1R, L2UIT
█
```

```
LinReg
y=a+bx
a=6439.954545
b=326.0818182
r^2=.9863642357
r=.9931587163
█
```

Now let's use the calculator to find a correlation. Unfortunately, the statistics package on your TI calculator does not automatically do that. Correlations are one of the most important things we might want to do, so here's how to fix that, once and for all.

- Hit **2nd CATALOG** (on the zero key). You now see a list of everything the calculator knows how to do. Impressive, huh?
- Scroll down until you find **DiagnosticOn**. Hit **ENTER**. Again. It should say **Done**.

Now and forevermore (or perhaps until you change batteries) your calculator will find correlations.

### Finding the Correlation

- *Always* check the conditions first. Look at the scatterplot for the Arizona State tuition data again. Does this association look linear? Are there outliers? This plot looks fine, but remember that correlation can be used to describe the strength of *linear* associations only, and outliers can distort the results. Eyeballing the scatterplot is an essential first step. (You should be getting used to checking on assumptions and conditions before jumping into a statistical procedure—it's always important.)
- Under the **STAT CALC** menu, select **8:LinReg(a+bx)** and hit **ENTER**.
- Now specify  $x$  and  $y$  by importing the names of your variables from the **LIST NAMES** menu. First name your  $x$ -variable followed by a comma, then your  $y$ -variable, creating the command

**LinReg(a+bx)L1R,L2UIT**

Wow! A lot of stuff happened. If you suspect all those other numbers are important, too, you'll really enjoy the next chapter. But for now, it's the value of  $r$  you care about. What does this correlation,  $r = 0.993$ , say about the trend in tuition costs?

## Correlation Properties

**AS** **Activity: Construct Scatterplots with a Given Correlation.** Try to make a scatterplot that has a given correlation. How close can you get?

### Height and Weight, Again

We could have measured the students' weights in stones. In the now outdated UK system of measures, a stone is a measure equal to 14 pounds. And we could have measured heights in hands. Hands are still commonly used to measure the heights of horses. A hand is 4 inches. But no matter what *units* we use to measure the two variables, the *correlation* stays the same.

### TI-*inspire*

**Correlation and Scatterplots.** See how the correlation changes as you drag data points around in a scatterplot.

Here's a useful list of facts about the correlation coefficient:

- ▶ The sign of a correlation coefficient gives the direction of the association.
- ▶ Correlation is always between  $-1$  and  $+1$ . Correlation *can* be exactly equal to  $-1.0$  or  $+1.0$ , but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
- ▶ Correlation treats  $x$  and  $y$  symmetrically. The correlation of  $x$  with  $y$  is the same as the correlation of  $y$  with  $x$ .
- ▶ Correlation has no units. This fact can be especially appropriate when the data's units are somewhat vague to begin with (IQ score, personality index, socialization, and so on). Correlation is sometimes given as a percentage, but you probably shouldn't do that because it suggests a percentage of *something*—and correlation, lacking units, has no “something” of which to be a percentage.
- ▶ Correlation is not affected by changes in the center or scale of either variable. Changing the units or baseline of either variable has no effect on the correlation coefficient. Correlation depends only on the  $z$ -scores, and they are unaffected by changes in center or scale.
- ▶ Correlation measures the strength of the *linear* association between the two variables. Variables can be strongly associated but still have a small correlation if the association isn't linear.
- ▶ Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small.

**How strong is strong?** You'll often see correlations characterized as “weak,” “moderate,” or “strong,” but be careful. There's no agreement on what those terms mean. The same numerical correlation might be strong in one context and weak in another. You might be thrilled to discover a correlation of  $0.7$  between the new summary of the economy you've come up with and stock market prices, but you'd consider it a design failure if you found a correlation of “only”  $0.7$  between two tests intended to measure the same skill. Deliberately vague terms like “weak,” “moderate,” or “strong” that describe a linear association can be useful additions to the numerical summary that correlation provides. But be sure to include the correlation and show a scatterplot, so others can judge for themselves.

### FOR EXAMPLE

#### Changing scales

**Recap:** We found a correlation of  $r = -0.879$  between hurricane wind speeds in knots and their central pressures in millibars.

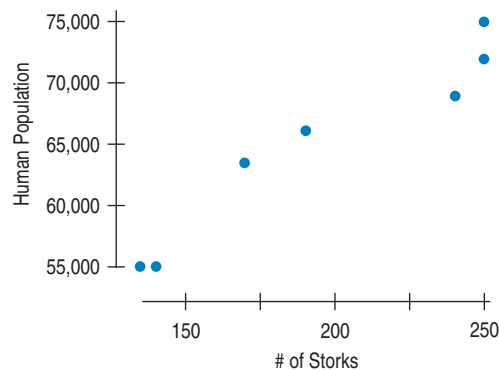
**Question:** Suppose we wanted to consider the wind speeds in miles per hour (1 mile per hour =  $0.869$  knots) and central pressures in inches of mercury (1 inch of mercury =  $33.86$  millibars). How would that conversion affect the conditions, the value of  $r$ , and our interpretation of the correlation coefficient?

Not at all! Correlation is based on standardized values ( $z$ -scores), so the conditions, the value of  $r$ , and the proper interpretation are all unaffected by changes in units.

## Warning: Correlation $\neq$ Causation

Whenever we have a strong correlation, it's tempting to try to explain it by imagining that the predictor variable has *caused* the response to change. Humans are like that; we tend to see causes and effects in everything.

Sometimes this tendency can be amusing. A scatterplot of the human population ( $y$ ) of Oldenburg, Germany, in the beginning of the 1930s plotted against the number of storks nesting in the town ( $x$ ) shows a tempting pattern.



**FIGURE 7.5**

The number of storks in Oldenburg, Germany, plotted against the population of the town for 7 years in the 1930s. The association is clear. How about the causation? (Ornithologishe Monatsberichte, 44, no. 2)

Anyone who has seen the beginning of the movie *Dumbo* remembers Mrs. Jumbo anxiously waiting for the stork to bring her new baby. Even though you know it's silly, you can't help but think for a minute that this plot shows that storks are the culprits. The two variables are obviously related to each other (the correlation is 0.97!), but that doesn't prove that storks bring babies.

It turns out that storks nest on house chimneys. More people means more houses, more nesting sites, and so more storks. The causation is actually in the *opposite* direction, but you can't tell from the scatterplot or correlation. You need additional information—not just the data—to determine the real mechanism.

A scatterplot of the damage (in dollars) caused to a house by fire would show a strong correlation with the number of firefighters at the scene. Surely the damage doesn't cause firefighters. And firefighters do seem to cause damage, spraying water all around and chopping holes. Does that mean we shouldn't call the fire department? Of course not. There is an underlying variable that leads to both more damage and more firefighters: the size of the blaze.

A hidden variable that stands behind a relationship and determines it by simultaneously affecting the other two variables is called a **lurking variable**. You can often debunk claims made about data by finding a lurking variable behind the scenes.

Scatterplots and correlation coefficients *never* prove causation. That's one reason it took so long for the U.S. Surgeon General to get warning labels on cigarettes. Although there was plenty of evidence that increased smoking was *associated* with increased levels of lung cancer, it took years to provide evidence that smoking actually *causes* lung cancer.

**Does cancer cause smoking?** Even if the correlation of two variables is due to a causal relationship, the correlation itself cannot tell us what causes what.

Sir Ronald Aylmer Fisher (1890–1962) was one of the greatest statisticians of the 20th century. Fisher testified in court (in testimony paid for by the tobacco companies) that a causal relationship might underlie the correlation of smoking and cancer:

“Is it possible, then, that lung cancer . . . is one of the causes of smoking cigarettes? I don't think it can be excluded . . . the pre-cancerous condition is one involving a certain amount of slight chronic inflammation . . .



A slight cause of irritation . . . is commonly accompanied by pulling out a cigarette, and getting a little compensation for life's minor ills in that way. And . . . is not unlikely to be associated with smoking more frequently."

Ironically, the proof that smoking indeed is the cause of many cancers came from experiments conducted following the principles of experiment design and analysis that Fisher himself developed—and that we'll see in Chapter 13.

## Correlation Tables

It is common in some fields to compute the correlations between every pair of variables in a collection of variables and arrange these correlations in a table. The rows and columns of the table name the variables, and the cells hold the correlations.

Correlation tables are compact and give a lot of summary information at a glance. They can be an efficient way to start to look at a large data set, but a dangerous one. By presenting all of these correlations without any checks for linearity and outliers, the correlation table risks showing truly small correlations that have been inflated by outliers, truly large correlations that are hidden by outliers, and correlations of any size that may be meaningless because the underlying form is not linear.

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

**Table 7.1**

A correlation table of data reported by *Forbes* magazine for large companies. From this table, can you be sure that the variables are linearly associated and free from outliers?

The diagonal cells of a correlation table always show correlations of exactly 1. (Can you see why?) Correlation tables are commonly offered by statistics packages on computers. These same packages often offer simple ways to make all the scatterplots that go with these correlations.

## Straightening Scatterplots

Correlation is a suitable measure of strength for straight relationships only. When a scatterplot shows a bent form that consistently increases or decreases, we can often straighten the form of the plot by re-expressing one or both variables.

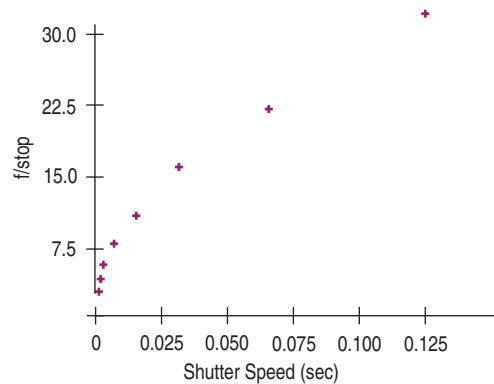
Some camera lenses have an adjustable aperture, the hole that lets the light in. The size of the aperture is expressed in a mysterious number called the *f/stop*. Each increase of one *f/stop* number corresponds to a halving of the light that is allowed to come through. The *f/stops* of one digital camera are

**f/stop:** 2.8 4 5.6 8 11 16 22 32

When you halve the shutter speed, you cut down the light, so you have to open the aperture one notch. We could experiment to find the best  $f$ /stop value for each shutter speed. A table of recommended shutter speeds and  $f$ /stops for a camera lists the relationship like this:

<b>Shutter speed:</b>	1/1000	1/500	1/250	1/125	1/60	1/30	1/15	1/8
<b><math>f</math>/stop:</b>	2.8	4	5.6	8	11	16	22	32

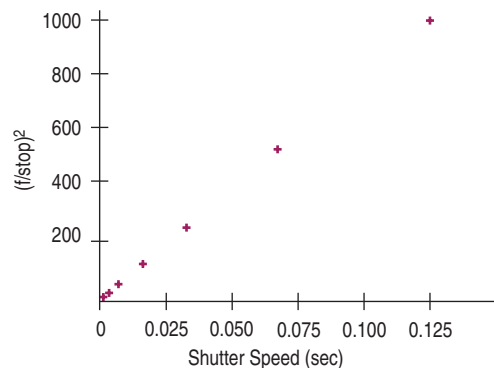
The correlation of these shutter speeds and  $f$ /stops is 0.979. That sounds pretty high. You might assume that there must be a strong linear relationship. But when we check the scatterplot (we *always* check the scatterplot), it shows that something is not quite right:



**FIGURE 7.6**

A scatterplot of  $f$ /stop vs. Shutter Speed shows a bent relationship.

We can see that the  $f$ /stop is not *linearly* related to the shutter speed. Can we find a transformation of  $f$ /stop that straightens out the line? What if we look at the *square* of the  $f$ /stop against the shutter speed?



**FIGURE 7.7**

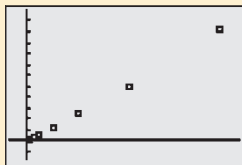
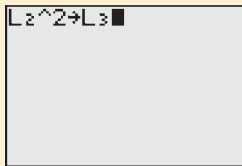
Re-expressing  $f$ /stop by squaring straightens the plot.

The second plot looks much more nearly straight. In fact, the correlation is now 0.998, but the increase in correlation is not important. (The original value of 0.979 should please almost anyone who sought a large correlation.) What is important is that the *form* of the plot is now straight, so the correlation is now an appropriate measure of association.<sup>10</sup>

We can often find transformations that straighten a scatterplot's form. Here, we found the square. Chapter 10 discusses simple ways to find a good re-expression.

<sup>10</sup> Sometimes we can do a "reality check" on our choice of re-expression. In this case, a bit of research reveals that  $f$ /stops are related to the diameter of the open shutter. Since the amount of light that enters is determined by the *area* of the open shutter, which is related to the diameter by squaring, the square re-expression seems reasonable. Not all re-expressions have such nice explanations, but it's a good idea to think about them.

## TI Tips



## Straightening a curve

Let's straighten the *f*/*stop* scatterplot with your calculator.

- Enter the data in two lists, *shutterspeed* in **L1** and *f/stop* in **L2**.
- Set up a **STAT PLOT** to create a scatterplot with **Xlist:L1** and **Ylist:L2**.
- Hit **ZoomStat**. See the curve?

We want to find the squares of all the *f*/*stops* and save those re-expressed values in another datalist. That's easy to do.

- Create the command to square all the values in **L2** and **STO**re those results in **L3**, then hit **ENTER**.

Now make the new scatterplot.

- Go back to **STAT PLOT** and change the setup. **Xlist** is still **L1**, but this time specify **Ylist:L3**.
- **ZoomStat** again.

You now see the straightened plot for these data. On deck: drawing the best line through those points!

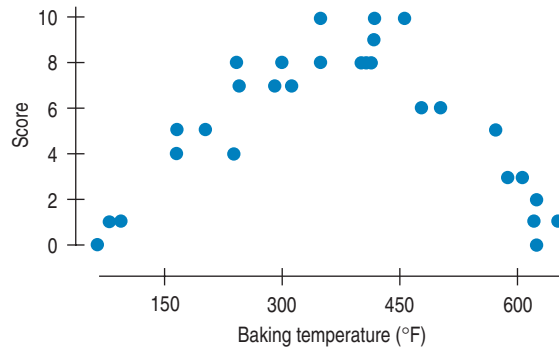
## WHAT CAN GO WRONG?

Did you know that there's a strong correlation between playing an instrument and drinking coffee? No? One reason might be that the statement doesn't make sense. Correlation is a statistic that's valid only for *quantitative* variables.

- ▶ **Don't say "correlation" when you mean "association."** How often have you heard the word "correlation"? Chances are pretty good that when you've heard the term, it's been misused. When people want to sound scientific, they often say "correlation" when talking about the relationship between two variables. It's one of the most widely misused Statistics terms, and given how often statistics are misused, that's saying a lot. One of the problems is that many people use the specific term *correlation* when they really mean the more general term *association*. "Association" is a deliberately vague term describing the relationship between two variables.

"Correlation" is a precise term that measures the strength and direction of the linear relationship between quantitative variables.

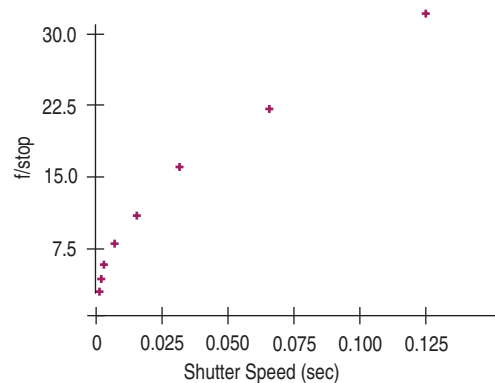
- ▶ **Don't correlate categorical variables.** People who misuse the term "correlation" to mean "association" often fail to notice whether the variables they discuss are quantitative. Be sure to check the **Quantitative Variables Condition**.
- ▶ **Don't confuse correlation with causation.** One of the most common mistakes people make in interpreting statistics occurs when they observe a high correlation between two variables and jump to the perhaps tempting conclusion that one thing must be causing the other. Scatterplots and correlations *never* demonstrate causation. At best, these statistical tools can only reveal an association between variables, and that's a far cry from establishing cause and effect. While it's true that some associations may be causal, the nature and direction of the causation can be very hard to establish, and there's always the risk of overlooking lurking variables.
- ▶ **Make sure the association is linear.** Not all associations between quantitative variables are linear. Correlation can miss even a strong nonlinear association. A student project evaluating the quality of brownies baked at different temperatures reports a correlation of  $-0.05$  between judges' scores and baking temperature. That seems to say there is no relationship—until we look at the scatterplot:



**FIGURE 7.8**  
The relationship between brownie taste Score and Baking Temperature is strong, but not at all linear.

There is a strong association, but the relationship is not linear. Don't forget to check the Straight Enough Condition.

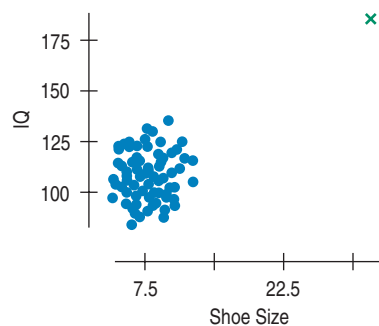
- ▶ **Don't assume the relationship is linear just because the correlation coefficient is high.** Recall that the correlation of  $f/\text{stops}$  and shutter speeds is 0.979 and yet the relationship is clearly not straight. Although the relationship must be straight for the correlation to be an appropriate measure, a high correlation is no guarantee of straightness. Nor is it safe to use correlation to judge the best re-expression. It's always important to look at the scatterplot.



**FIGURE 7.9**  
A scatterplot of  $f/\text{stop}$  vs. Shutter Speed shows a bent relationship even though the correlation is  $r = 0.979$ .



- ▶ **Beware of outliers.** You can't interpret a correlation coefficient safely without a background check for outliers. Here's a silly example:  
The relationship between IQ and shoe size among comedians shows a surprisingly strong positive correlation of 0.50. To check assumptions, we look at the scatterplot:

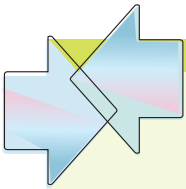


**FIGURE 7.10**  
A scatterplot of IQ vs. Shoe Size. From this "study," what is the relationship between the two? The correlation is 0.50. Who does that point (the green x) in the upper right-hand corner belong to?

The outlier is Bozo the Clown, known for his large shoes, and widely acknowledged to be a comic "genius." Without Bozo, the correlation is near zero.

Even a single outlier can dominate the correlation value. That's why you need to check the Outlier Condition.



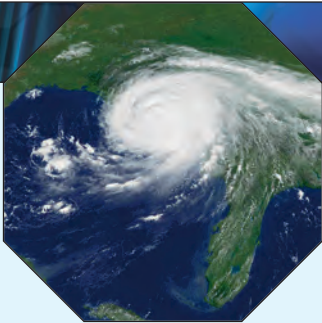


## CONNECTIONS

Scatterplots are the basic tool for examining the relationship between two quantitative variables. We start with a picture when we want to understand the distribution of a single variable, and we always make a scatterplot to begin to understand the relationship between two quantitative variables.

We used  $z$ -scores as a way to measure the statistical distance of data values from their means. Now we've seen the  $z$ -scores of  $x$  and  $y$  working together to build the correlation coefficient. Correlation is a summary statistic like the mean and standard deviation—only it summarizes the strength of a linear relationship. And we interpret it as we did  $z$ -scores, using the standard deviations as our rulers in both  $x$  and  $y$ .

## WHAT HAVE WE LEARNED?



In recent chapters we learned how to listen to the story told by data from a single variable. Now we've turned our attention to the more complicated (and more interesting) story we can discover in the association between two quantitative variables.

We've learned to begin our investigation by looking at a scatterplot. We're interested in the *direction* of the association, the *form* it takes, and its *strength*.

We've learned that, although not every relationship is linear, when the scatterplot is straight enough, the *correlation coefficient* is a useful numerical summary.

- ▶ The sign of the correlation tells us the direction of the association.
- ▶ The magnitude of the correlation tells us the *strength* of a linear association. Strong associations have correlations near  $-1$  or  $+1$  and very weak associations near  $0$ .
- ▶ Correlation has no units, so shifting or scaling the data, standardizing, or even swapping the variables has no effect on the numerical value.

Once again we've learned that doing Statistics right means we have to *Think* about whether our choice of methods is appropriate.

- ▶ The correlation coefficient is appropriate only if the underlying relationship is linear.
- ▶ We'll check the **Straight Enough Condition** by looking at a scatterplot.
- ▶ And, as always, we'll watch out for outliers!

Finally, we've learned not to make the mistake of assuming that a high correlation or strong association is evidence of a cause-and-effect relationship. Beware of lurking variables!

**A S** **Simulation: Correlation, Center, and Scale.** If you have any lingering doubts that shifting and rescaling the data won't change the correlation, watch nothing happen right before your eyes!

## Terms

### Scatterplots

147. A scatterplot shows the relationship between two quantitative variables measured on the same cases.

### Association

- ▶ 147. **Direction:** A positive direction or association means that, in general, as one variable increases, so does the other. When increases in one variable generally correspond to decreases in the other, the association is negative.
- ▶ 147. **Form:** The form we care about most is straight, but you should certainly describe other patterns you see in scatterplots.
- ▶ 148. **Strength:** A scatterplot is said to show a strong association if there is little scatter around the underlying relationship.

### Outlier

148. A point that does not fit the overall pattern seen in the scatterplot.

Response variable,  
Explanatory variable,  
x-variable, y-variable  
Correlation Coefficient

149. In a scatterplot, you must choose a role for each variable. Assign to the  $y$ -axis the response variable that you hope to predict or explain. Assign to the  $x$ -axis the explanatory or predictor variable that accounts for, explains, predicts, or is otherwise responsible for the  $y$ -variable.

152. The correlation coefficient is a numerical measure of the direction and strength of a linear association.

$$r = \frac{\sum z_x z_y}{n - 1}$$

Lurking variable

157. A variable other than  $x$  and  $y$  that simultaneously affects both variables, accounting for the correlation between the two.

## Skills

THINK

- ▶ Recognize when interest in the pattern of a possible relationship between two quantitative variables suggests making a scatterplot.
- ▶ Know how to identify the roles of the variables and that you should place the response variable on the  $y$ -axis and the explanatory variable on the  $x$ -axis.
- ▶ Know the conditions for correlation and how to check them.
- ▶ Know that correlations are between  $-1$  and  $+1$ , and that each extreme indicates a perfect linear association.
- ▶ Understand how the magnitude of the correlation reflects the strength of a linear association as viewed in a scatterplot.
- ▶ Know that correlation has no units.
- ▶ Know that the correlation coefficient is not changed by changing the center or scale of either variable.

SHOW

- ▶ Understand that causation cannot be demonstrated by a scatterplot or correlation.
- ▶ Know how to make a scatterplot by hand (for a small set of data) or with technology.
- ▶ Know how to compute the correlation of two variables.

TELL

- ▶ Know how to read a correlation table produced by a statistics program.
- ▶ Be able to describe the direction, form, and strength of a scatterplot.
- ▶ Be prepared to identify and describe points that deviate from the overall pattern.
- ▶ Be able to use correlation as part of the description of a scatterplot.
- ▶ Be alert to misinterpretations of correlation.
- ▶ Understand that finding a correlation between two variables does not indicate a causal relationship between them. Beware the dangers of suggesting causal relationships when describing correlations.

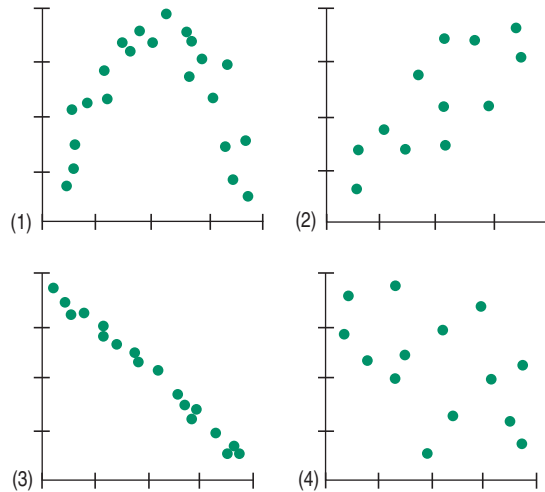
## SCATTERPLOTS AND CORRELATION ON THE COMPUTER

Statistics packages generally make it easy to look at a scatterplot to check whether the correlation is appropriate. Some packages make this easier than others.

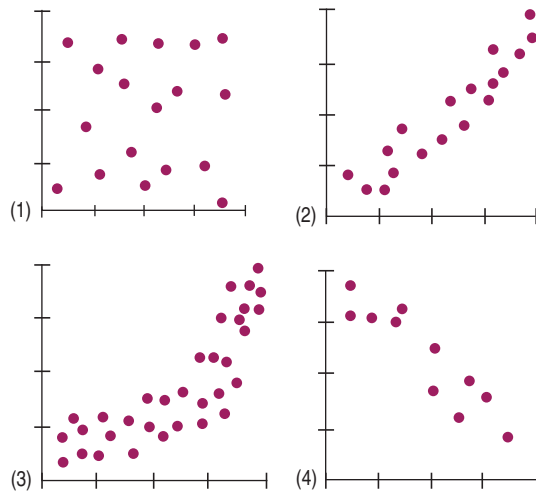
Many packages allow you to modify or enhance a scatterplot, altering the axis labels, the axis numbering, the plot symbols, or the colors used. Some options, such as color and symbol choice, can be used to display additional information on the scatterplot.

## EXERCISES

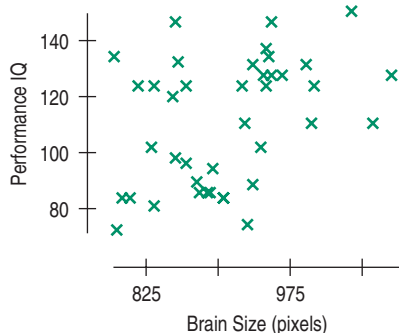
- Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
  - Apples: weight in grams, weight in ounces
  - Apples: circumference (inches), weight (ounces)
  - College freshmen: shoe size, grade point average
  - Gasoline: number of miles you drove since filling up, gallons remaining in your tank
- Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
  - T-shirts at a store: price each, number sold
  - Scuba diving: depth, water pressure
  - Scuba diving: depth, visibility
  - All elementary school students: weight, score on a reading test
- Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
  - When climbing mountains: altitude, temperature
  - For each week: ice cream cone sales, air-conditioner sales
  - People: age, grip strength
  - Drivers: blood alcohol level, reaction time
- Association.** Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.
  - Long-distance calls: time (minutes), cost
  - Lightning strikes: distance from lightning, time delay of the thunder
  - A streetlight: its apparent brightness, your distance from it
  - Cars: weight of car, age of owner
- Scatterplots.** Which of the scatterplots at the top of the next column show
  - little or no association?
  - a negative association?
  - a linear association?
  - a moderately strong association?
  - a very strong association?



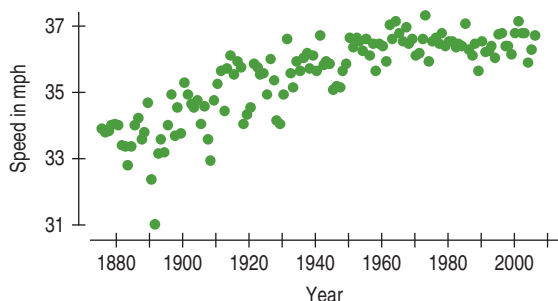
- Scatterplots.** Which of the scatterplots below show
  - little or no association?
  - a negative association?
  - a linear association?
  - a moderately strong association?
  - a very strong association?



- Performance IQ scores vs. brain size.** A study examined brain size (measured as pixels counted in a digitized magnetic resonance image [MRI] of a cross section of the brain) and IQ (4 Performance scales of the Weschler IQ test) for college students. The scatterplot shows the Performance IQ scores vs. the brain size. Comment on the association between brain size and IQ as seen in the scatterplot on the next page.

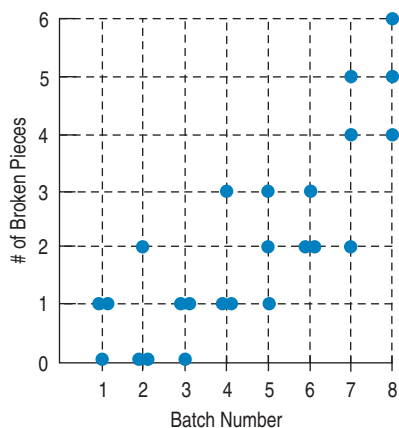


- 8. Kentucky Derby 2006.** The fastest horse in Kentucky Derby history was Secretariat in 1973. The scatterplot shows speed (in miles per hour) of the winning horses each year.



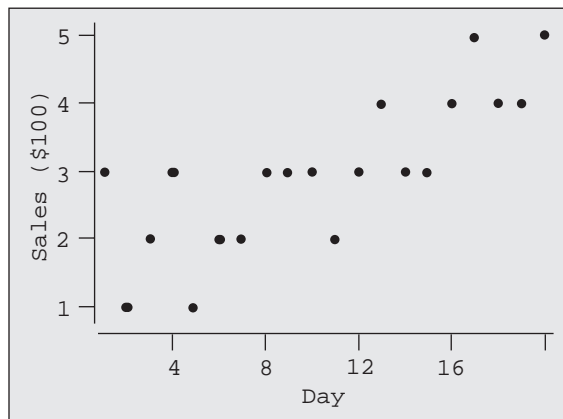
What do you see? In most sporting events, performances have improved and continue to improve, so surely we anticipate a positive direction. But what of the form? Has the performance increased at the same rate throughout the last 130 years?

- 9. Firing pottery.** A ceramics factory can fire eight large batches of pottery a day. Sometimes a few of the pieces break in the process. In order to understand the problem better, the factory records the number of broken pieces in each batch for 3 days and then creates the scatterplot shown.

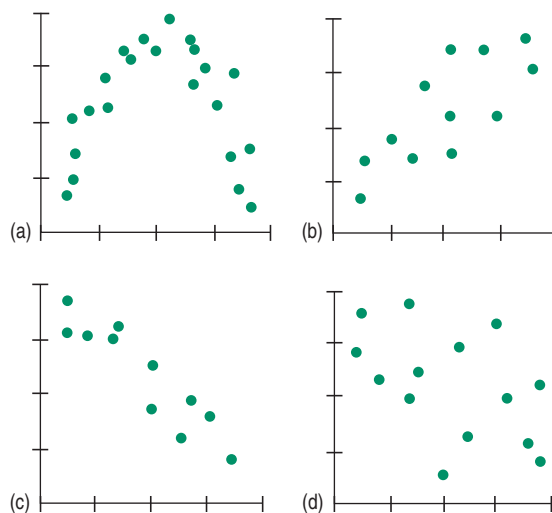


- Make a histogram showing the distribution of the number of broken pieces in the 24 batches of pottery examined.
- Describe the distribution as shown in the histogram. What feature of the problem is more apparent in the histogram than in the scatterplot?
- What aspect of the company's problem is more apparent in the scatterplot?

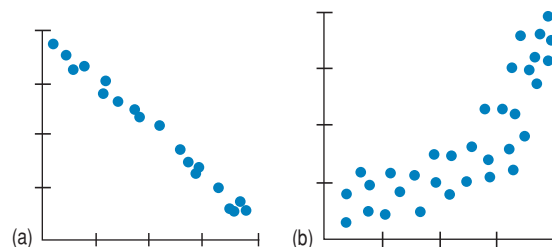
- 10. Coffee sales.** Owners of a new coffee shop tracked sales for the first 20 days and displayed the data in a scatterplot (by day).



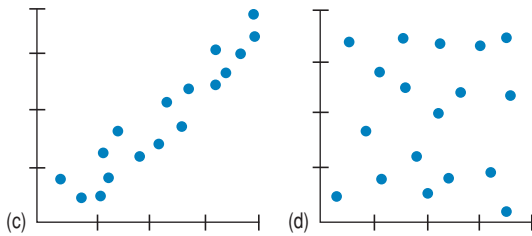
- Make a histogram of the daily sales since the shop has been in business.
  - State one fact that is obvious from the scatterplot, but not from the histogram.
  - State one fact that is obvious from the histogram, but not from the scatterplot.
- 11. Matching.** Here are several scatterplots. The calculated correlations are  $-0.923$ ,  $-0.487$ ,  $0.006$ , and  $0.777$ . Which is which?



- 12. Matching.** Here and on the next page are several scatterplots. The calculated correlations are  $-0.977$ ,  $-0.021$ ,  $0.736$ , and  $0.951$ . Which is which?

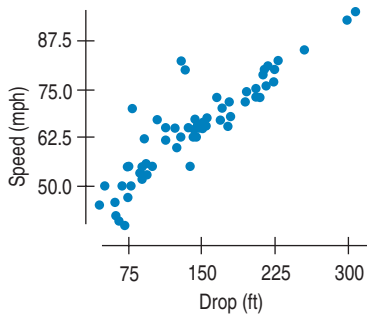




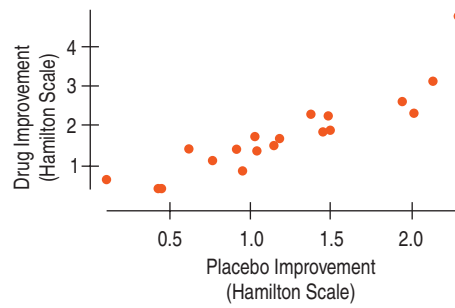


13. **Politics.** A candidate for office claims that “there is a correlation between television watching and crime.” Criticize this statement on statistical grounds.
14. **Car thefts.** The National Insurance Crime Bureau reports that Honda Accords, Honda Civics, and Toyota Camrys are the cars most frequently reported stolen, while Ford Tauruses, Pontiac Vibes, and Buick LeSabres are stolen least often. Is it reasonable to say that there’s a correlation between the type of car you own and the risk that it will be stolen?

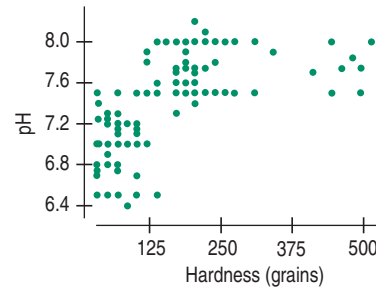
- T 15. **Roller coasters.** Roller coasters get all their speed by dropping down a steep initial incline, so it makes sense that the height of that drop might be related to the speed of the coaster. Here’s a scatterplot of top *Speed* and largest *Drop* for 75 roller coasters around the world.



- a) Does the scatterplot indicate that it is appropriate to calculate the correlation? Explain.
- b) In fact, the correlation of *Speed* and *Drop* is 0.91. Describe the association.
- T 16. **Antidepressants.** A study compared the effectiveness of several antidepressants by examining the experiments in which they had passed the FDA requirements. Each of those experiments compared the active drug with a placebo, an inert pill given to some of the subjects. In each experiment some patients treated with the placebo had improved, a phenomenon called the *placebo effect*. Patients’ depression levels were evaluated on the Hamilton Depression Rating Scale, where larger numbers indicate greater improvement. (The Hamilton scale is a widely accepted standard that was used in each of the independently run studies.) The scatterplot at the top of the next column compares mean improvement levels for the antidepressants and placebos for several experiments.

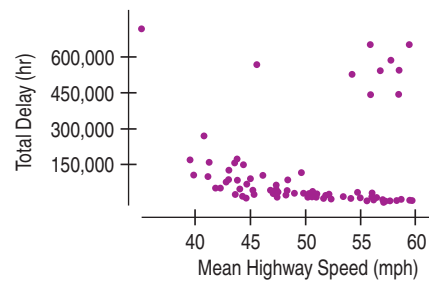


- a) Is it appropriate to calculate the correlation? Explain.
- b) The correlation is 0.898. Explain what we have learned about the results of these experiments.
- T 17. **Hard water.** In a study of streams in the Adirondack Mountains, the following relationship was found between the water’s pH and its hardness (measured in grains):



Is it appropriate to summarize the strength of association with a correlation? Explain.

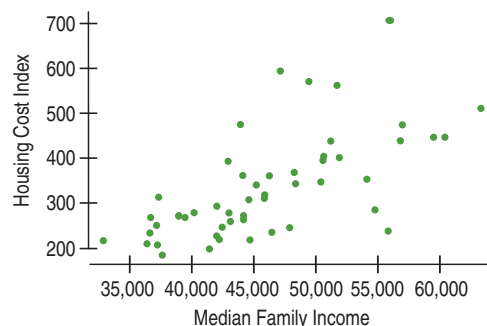
18. **Traffic headaches.** A study of traffic delays in 68 U.S. cities found the following relationship between total delays (in total hours lost) and mean highway speed:



Is it appropriate to summarize the strength of association with a correlation? Explain.

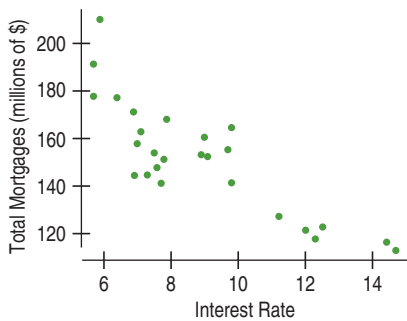
19. **Cold nights.** Is there an association between time of year and the nighttime temperature in North Dakota? A researcher assigned the numbers 1–365 to the days January 1–December 31 and recorded the temperature at 2:00 a.m. for each. What might you expect the correlation between *DayNumber* and *Temperature* to be? Explain.

20. **Association.** A researcher investigating the association between two variables collected some data and was surprised when he calculated the correlation. He had expected to find a fairly strong association, yet the correlation was near 0. Discouraged, he didn't bother making a scatterplot. Explain to him how the scatterplot could still reveal the strong association he anticipated.
21. **Prediction units.** The errors in predicting hurricane tracks (examined in this chapter) were given in nautical miles. An ordinary mile is 0.86898 nautical miles. Most people living on the Gulf Coast of the United States would prefer to know the prediction errors in miles rather than nautical miles. Explain why converting the errors to miles would not change the correlation between *Prediction Error* and *Year*.
22. **More predictions.** Hurricane Katrina's hurricane force winds extended 120 miles from its center. Katrina was a big storm, and that affects how we think about the prediction errors. Suppose we add 120 miles to each error to get an idea of how far from the predicted track we might still find damaging winds. Explain what would happen to the correlation between *Prediction Error* and *Year*, and why.
23. **Correlation errors.** Your Economics instructor assigns your class to investigate factors associated with the gross domestic product (*GDP*) of nations. Each student examines a different factor (such as *Life Expectancy*, *Literacy Rate*, etc.) for a few countries and reports to the class. Apparently, some of your classmates do not understand Statistics very well because you know several of their conclusions are incorrect. Explain the mistakes in their statements below.
- "My very low correlation of  $-0.772$  shows that there is almost no association between *GDP* and *Infant Mortality Rate*."
  - "There was a correlation of  $0.44$  between *GDP* and *Continent*."
24. **More correlation errors.** Students in the Economics class discussed in Exercise 23 also wrote these conclusions. Explain the mistakes they made.
- "There was a very strong correlation of  $1.22$  between *Life Expectancy* and *GDP*."
  - "The correlation between *Literacy Rate* and *GDP* was  $0.83$ . This shows that countries wanting to increase their standard of living should invest heavily in education."
25. **Height and reading.** A researcher studies children in elementary school and finds a strong positive linear association between height and reading scores.
- Does this mean that taller children are generally better readers?
  - What might explain the strong correlation?
26. **Cellular telephones and life expectancy.** A survey of the world's nations in 2004 shows a strong positive correlation between percentage of the country using cell phones and life expectancy in years at birth.
- Does this mean that cell phones are good for your health?
  - What might explain the strong correlation?
27. **Correlation conclusions I.** The correlation between *Age* and *Income* as measured on 100 people is  $r = 0.75$ . Explain whether or not each of these possible conclusions is justified:
- When *Age* increases, *Income* increases as well.
  - The form of the relationship between *Age* and *Income* is straight.
  - There are no outliers in the scatterplot of *Income* vs. *Age*.
  - Whether we measure *Age* in years or months, the correlation will still be  $0.75$ .
28. **Correlation conclusions II.** The correlation between *Fuel Efficiency* (as measured by miles per gallon) and *Price* of 150 cars at a large dealership is  $r = -0.34$ . Explain whether or not each of these possible conclusions is justified:
- The more you pay, the lower the fuel efficiency of your car will be.
  - The form of the relationship between *Fuel Efficiency* and *Price* is moderately straight.
  - There are several outliers that explain the low correlation.
  - If we measure *Fuel Efficiency* in kilometers per liter instead of miles per gallon, the correlation will increase.
29. **Baldness and heart disease.** Medical researchers followed 1435 middle-aged men for a period of 5 years, measuring the amount of *Baldness* present (none = 1, little = 2, some = 3, much = 4, extreme = 5) and presence of *Heart Disease* (No = 0, Yes = 1). They found a correlation of  $0.089$  between the two variables. Comment on their conclusion that this shows that baldness is not a possible cause of heart disease.
30. **Sample survey.** A polling organization is checking its database to see if the two data sources it used sampled the same zip codes. The variable *Datasource* = 1 if the data source is MetroMedia, 2 if the data source is DataQwest, and 3 if it's RollingPoll. The organization finds that the correlation between five-digit zip code and *Datasource* is  $-0.0229$ . It concludes that the correlation is low enough to state that there is no dependency between *Zip Code* and *Source of Data*. Comment.
31. **Income and housing.** The Office of Federal Housing Enterprise Oversight ([www.ofheo.gov](http://www.ofheo.gov)) collects data on various aspects of housing costs around the United States. Here is a scatterplot of the *Housing Cost Index* versus the *Median Family Income* for each of the 50 states. The correlation is  $0.65$ .



- a) Describe the relationship between the *Housing Cost Index* and the *Median Family Income* by state.
- b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
- c) If we had measured *Median Family Income* in thousands of dollars instead of dollars, how would the correlation change?
- d) Washington, DC, has a Housing Cost Index of 548 and a median income of about \$45,000. If we were to include DC in the data set, how would that affect the correlation coefficient?
- e) Do these data provide proof that by raising the median income in a state, the Housing Cost Index will rise as a result? Explain.

**T 32. Interest rates and mortgages.** Since 1980, average mortgage interest rates have fluctuated from a low of under 6% to a high of over 14%. Is there a relationship between the amount of money people borrow and the interest rate that's offered? Here is a scatterplot of *Total Mortgages* in the United States (in millions of 2005 dollars) versus *Interest Rate* at various times over the past 26 years. The correlation is  $-0.84$ .



- a) Describe the relationship between *Total Mortgages* and *Interest Rate*.
- b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
- c) If we were to measure *Total Mortgages* in thousands of dollars instead of millions of dollars, how would the correlation coefficient change?
- d) Suppose in another year, interest rates were 11% and mortgages totaled \$250 million. How would including that year with these data affect the correlation coefficient?
- e) Do these data provide proof that if mortgage rates are lowered, people will take out more mortgages? Explain.

**T 33. Fuel economy 2007.** Here are advertised horsepower ratings and expected gas mileage for several 2007 vehicles. (<http://www.kbb.com/KBB/ReviewsAndRatings>)

Vehicle	Horsepower	Highway Gas Mileage (mpg)
Audi A4	200	32
BMW 328	230	30
Buick LaCrosse	200	30
Chevy Cobalt	148	32
Chevy TrailBlazer	291	22
Ford Expedition	300	20
GMC Yukon	295	21
Honda Civic	140	40
Honda Accord	166	34
Hyundai Elantra	138	36
Lexus IS 350	306	28
Lincoln Navigator	300	18
Mazda Tribute	212	25
Toyota Camry	158	34
Volkswagen Beetle	150	30

- a) Make a scatterplot for these data.
  - b) Describe the direction, form, and strength of the plot.
  - c) Find the correlation between horsepower and miles per gallon.
  - d) Write a few sentences telling what the plot says about fuel economy.
- 34. Drug abuse.** A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the table.

Country	Percent Who Have Used	
	Marijuana	Other Drugs
Czech Rep.	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
USA	34	24

- a) Create a scatterplot.
- b) What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs?
- c) Write a brief description of the association.
- d) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs? Explain.

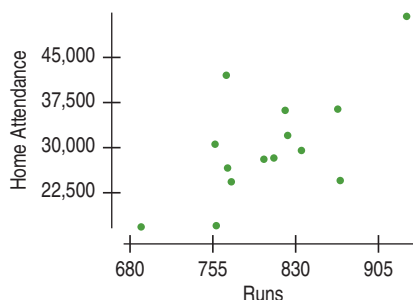
- T 35. Burgers.** Fast food is often considered unhealthy because much of it is high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers. Analyze the association between fat content and sodium.

Fat (g)	19	31	34	35	39	39	43
Sodium (mg)	920	1500	1310	860	1180	940	1260

- T 36. Burgers II.** In the previous exercise you analyzed the association between the amounts of fat and sodium in fast food hamburgers. What about fat and calories? Here are data for the same burgers:

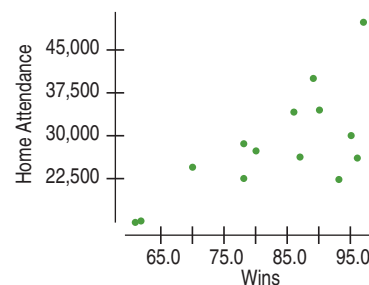
Fat (g)	19	31	34	35	39	39	43
Calories	410	580	590	570	640	680	660

- T 37. Attendance 2006.** American League baseball games are played under the designated hitter rule, meaning that pitchers, often weak hitters, do not come to bat. Baseball owners believe that the designated hitter rule means more runs scored, which in turn means higher attendance. Is there evidence that more fans attend games if the teams score more runs? Data collected from American League games during the 2006 season indicate a correlation of 0.667 between runs scored and the number of people at the game. (<http://mlb.mlb.com>)



- a) Does the scatterplot indicate that it's appropriate to calculate a correlation? Explain.
- b) Describe the association between attendance and runs scored.
- c) Does this association prove that the owners are right that more fans will come to games if the teams score more runs?
- T 38. Second inning 2006.** Perhaps fans are just more interested in teams that win. The displays below are based on American League teams for the 2006 season. (<http://espn.go.com>) Are the teams that win necessarily those which score the most runs?

CORRELATION			
	Wins	Runs	Attend
Wins	1.000		
Runs	0.605	1.000	
Attend	0.697	0.667	1.000



- a) Do winning teams generally enjoy greater attendance at their home games? Describe the association.
- b) Is attendance more strongly associated with winning or scoring runs? Explain.
- c) How strongly is scoring more runs associated with winning more games?
- T 39. Thrills.** People who responded to a July 2004 Discovery Channel poll named the 10 best roller coasters in the United States. The table below shows the length of the initial drop (in feet) and the duration of the ride (in seconds). What do these data indicate about the height of a roller coaster and the length of the ride you can expect?

Roller Coaster	State	Drop (ft)	Duration (sec)
Incredible Hulk	FL	105	135
Millennium Force	OH	300	105
Goliath	CA	255	180
Nitro	NJ	215	240
Magnum XL-2000	OH	195	120
The Beast	OH	141	65
Son of Beast	OH	214	140
Thunderbolt	PA	95	90
Ghost Rider	CA	108	160
Raven	IN	86	90

- T 40. Vehicle weights.** The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed "weight-in-motion" scale. To see if the new device was accurate, they conducted a calibration test. They weighed several stopped trucks (static weight) and assumed that this weight was correct. Then they weighed the trucks again while they were moving to see how well the new scale could estimate the actual weight. Their data are given in the table on the next page.

WEIGHTS (1000s OF LBS)	
Weight-in-Motion	Static Weight
26.0	27.9
29.9	29.1
39.5	38.0
25.1	27.0
31.6	30.3
36.2	34.5
25.1	27.8
31.0	29.6
35.6	33.1
40.2	35.5

- a) Make a scatterplot for these data.  
 b) Describe the direction, form, and strength of the plot.  
 c) Write a few sentences telling what the plot says about the data. (*Note:* The sentences should be about weighing trucks, not about scatterplots.)  
 d) Find the correlation.  
 e) If the trucks were weighed in kilograms, how would this change the correlation? (1 kilogram = 2.2 pounds)  
 f) Do any points deviate from the overall pattern? What does the plot say about a possible recalibration of the weight-in-motion scale?
41. **Planets (more or less).** On August 24, 2006, the International Astronomical Union voted that Pluto is not a planet. Some members of the public have been reluctant to accept that decision. Let's look at some of the data. (We'll see more in the next chapter.) Is there any pattern to the locations of the planets? The table shows the average distance of each of the traditional nine planets from the sun.

Planet	Position Number	Distance from Sun (million miles)
Mercury	1	36
Venus	2	67
Earth	3	93
Mars	4	142
Jupiter	5	484
Saturn	6	887
Uranus	7	1784
Neptune	8	2796
Pluto	9	3666

- a) Make a scatterplot and describe the association. (Remember: direction, form, and strength!)  
 b) Why would you not want to talk about the correlation between a planet's *Position* and *Distance* from the sun?  
 c) Make a scatterplot showing the logarithm of *Distance* vs. *Position*. What is better about this scatterplot?

42. **Flights.** The number of flights by U.S. Airlines has grown rapidly. Here are the number of flights flown in each year from 1995 to 2005.
- Find the correlation of *Flights* with *Year*.
  - Make a scatterplot and describe the trend.
  - Note two reasons that the correlation you found in (a) is not a suitable summary of the strength of the association. Can you account for these violations of the conditions?

Year	Flights
1995	5,327,435
1996	5,351,983
1997	5,411,843
1998	5,384,721
1999	5,527,884
2000	5,683,047
2001	5,967,780
2002	5,271,359
2003	6,488,539
2004	7,129,270
2005	7,140,596



### JUST CHECKING Answers

- We know the scores are quantitative. We should check to see if the Straight Enough Condition and the Outlier Condition are satisfied by looking at a scatterplot of the two scores.
- It won't change.
- It won't change.
- They are likely to have done poorly. The positive correlation means that low scores on Exam 1 are associated with low scores on Exam 2 (and similarly for high scores).
- No. The general association is positive, but individual performances may vary.

# Linear Regression



**WHO** Items on the Burger King menu

**WHAT** Protein content and total fat content

**UNITS** Grams of protein  
Grams of fat

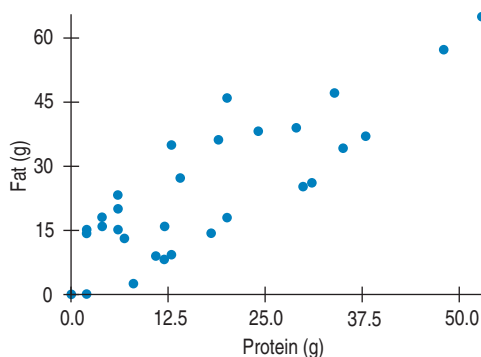
**HOW** Supplied by BK on request or at their Web site

**A S** **Video: Manatees and Motorboats.** Are motorboats killing more manatees in Florida? Here's the story on video.

**A S** **Activity: Linear Equations.** For a quick review of linear equations, view this activity and play with the interactive tool.

The Whopper™ has been Burger King's signature sandwich since 1957. One Double Whopper with cheese provides 53 grams of protein—all the protein you need in a day. It also supplies 1020 calories and 65 grams of fat. The Daily Value (based on a 2000-calorie diet) for fat is 65 grams. So after a Double Whopper you'll want the rest of your calories that day to be fat-free.<sup>1</sup>

Of course, the Whopper isn't the only item Burger King sells. How are fat and protein related on the entire BK menu? The scatterplot of the *Fat* (in grams) versus the *Protein* (in grams) for foods sold at Burger King shows a positive, moderately strong, linear relationship.



**FIGURE 8.1**

*Total Fat versus Protein for 30 items on the BK menu. The Double Whopper is in the upper right corner. It's extreme, but is it out of line?*

If you want 25 grams of protein in your lunch, how much fat should you expect to consume at Burger King? The correlation between *Fat* and *Protein* is 0.83, a sign that the linear association seen in the scatterplot is fairly strong. But *strength* of the relationship is only part of the picture. The correlation says, "The linear association between these two variables is fairly strong," but it doesn't tell us *what the line is*.

<sup>1</sup> Sorry about the fries.

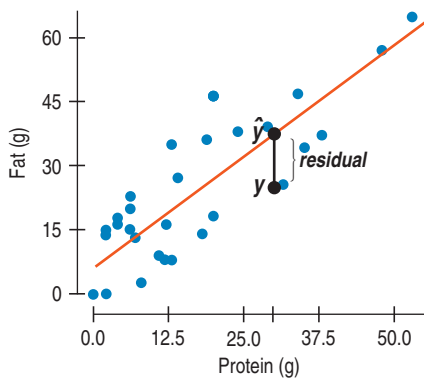
“Statisticians, like artists, have the bad habit of falling in love with their models.”

—George Box, famous statistician

### Activity: Residuals.

Residuals are the basis for fitting lines to scatterplots. See how they work.

## Residuals



$\text{residual} = \text{observed value} - \text{predicted value}$

A *negative* residual means the predicted value is too big—an overestimate. And a *positive* residual shows that the model makes an underestimate. These may seem backwards until you think about them.

Now we can say more. We can **model** the relationship with a line and give its **equation**. The equation will let us predict the fat content for any Burger King food, given its amount of protein.

We met our first model in Chapter 6. We saw there that we can specify a Normal model with two parameters: its mean ( $\mu$ ) and standard deviation ( $\sigma$ ).

For the Burger King foods, we’d choose a linear model to describe the relationship between *Protein* and *Fat*. **The linear model is just an equation of a straight line through the data.** Of course, no line can go through all the points, but a linear model can summarize the general pattern with only a couple of parameters. Like all models of the real world, the line will be wrong—wrong in the sense that it can’t match reality *exactly*. But it can help us understand how the variables are associated.

Not only can’t we draw a line through all the points, the best line might not even hit *any* of the points. Then how can it be the “best” line? We want to find the line that somehow comes *closer* to all the points than any other line. Some of the points will be above the line and some below. For example, the line might suggest that a BK Broiler chicken sandwich with 30 grams of protein should have 36 grams of fat when, in fact, it actually has only 25 grams of fat. We call the estimate made from a model the **predicted value**, and write it as  $\hat{y}$  (called *y-hat*) to distinguish it from the true value  $y$  (called, uh,  $y$ ). The difference between the observed value and its associated predicted value is called the **residual**. The residual value tells us how far off the model’s prediction is at that point. The BK Broiler chicken residual would be  $y - \hat{y} = 25 - 36 = -11$  g of fat.

To find the residuals, we always subtract the predicted value from the observed one. The negative residual tells us that the actual fat content of the BK Broiler chicken is about 11 grams *less* than the model predicts for a typical Burger King menu item with 30 grams of protein.

Our challenge now is how to find the right line.

## “Best Fit” Means Least Squares

**Activity: The Least Squares Criterion.** Does your sense of “best fit” look like the least squares line?

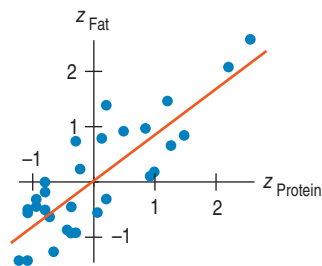
### Who’s on First

In 1805, Legendre was the first to publish the “least squares” solution to the problem of fitting a line to data when the points don’t all fall exactly on the line. The main challenge was how to distribute the errors “fairly.” After considerable thought, he decided to minimize the sum of the squares of what we now call the residuals. When Legendre published his paper, though, Gauss claimed he had been using the method since 1795. Gauss later referred to the “least squares” solution as “our method” (*principium nostrum*), which certainly didn’t help his relationship with Legendre.

When we draw a line through a scatterplot, some residuals are positive and some negative. We can’t assess how well the line fits by adding up all the residuals—the positive and negative ones would just cancel each other out. We faced the same issue when we calculated a standard deviation to measure spread. And we deal with it the same way here: by squaring the residuals. Squaring makes them all positive. Now we can add them up. Squaring also emphasizes the large residuals. After all, points near the line are consistent with the model, but we’re more concerned about points far from the line. When we add all the squared residuals together, that sum indicates how well the line we drew fits the data—the smaller the sum, the better the fit. A different line will produce a different sum, maybe bigger, maybe smaller. **The line of best fit is the line for which the sum of the squared residuals is smallest, the least squares line.**

TI-*n*spire

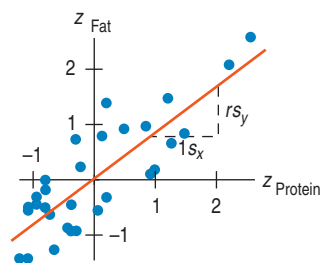
**Least squares.** Try to minimize the sum of areas of residual squares as you drag a line across a scatterplot.



**FIGURE 8.2**  
The Burger King scatterplot in z-scores.

**NOTATION ALERT:**

“Putting a hat on it” is standard Statistics notation to indicate that something has been predicted by a model. Whenever you see a hat over a variable name or symbol, you can assume it is the predicted version of that variable or symbol (and look around for the model).



**FIGURE 8.3**  
Standardized fat vs. standardized protein with the regression line. Each one standard deviation change in protein results in a predicted change of  $r$  standard deviations in fat.

You might think that finding this line would be pretty hard. Surprisingly, it's not, although it was an exciting mathematical discovery when Legendre published it in 1805 (see margin note on previous page).

## Correlation and the Line

If you suspect that what we know about correlation can lead us to the equation of the linear model, you're headed in the right direction. It turns out that it's not a very big step. In Chapter 7 we learned a lot about how correlation worked by looking at a scatterplot of the standardized variables. Here's a scatterplot of  $z_y$  (standardized *Fat*) vs.  $z_x$  (standardized *Protein*).

What line would you choose to model the relationship of the standardized values? Let's start at the center of the scatterplot. How much protein and fat does a *typical* Burger King food item provide? If it has average protein content,  $\bar{x}$ , what about its fat content? If you guessed that its fat content should be about average,  $\bar{y}$ , as well, then you've discovered the first property of the line we're looking for. The line must go through the point  $(\bar{x}, \bar{y})$ . In the plot of z-scores, then, the line passes through the origin  $(0, 0)$ .

You might recall that the equation for a line that passes through the origin can be written with just a slope and no intercept:

$$y = mx.$$

The coordinates of our standardized points aren't written  $(x, y)$ ; their coordinates are z-scores:  $(z_x, z_y)$ . We'll need to change our equation to show that. And we'll need to indicate that the point on the line corresponding to a particular  $z_x$  is  $\hat{z}_y$ , the model's estimate of the actual value of  $z_y$ . So our equation becomes

$$\hat{z}_y = mz_x.$$

Many lines with different slopes pass through the origin. Which one fits our data the best? That is, which slope determines the line that minimizes the sum of the squared residuals? It turns out that the best choice for  $m$  is the correlation coefficient itself,  $r$ ! (You must really wonder where that stunning assertion comes from. Check the Math Box.)

Wow! This line has an equation that's about as simple as we could possibly hope for:

$$\hat{z}_y = rz_x.$$

Great. It's simple, but what does it tell us? It says that in moving one standard deviation from the mean in  $x$ , we can expect to move about  $r$  standard deviations away from the mean in  $y$ . Now that we're thinking about least squares lines, the correlation is more than just a vague measure of strength of association. It's a great way to think about what the model tells us.

Let's be more specific. For the sandwiches, the correlation is 0.83. If we standardize both protein and fat, we can write

$$\hat{z}_{Fat} = 0.83z_{Protein}.$$

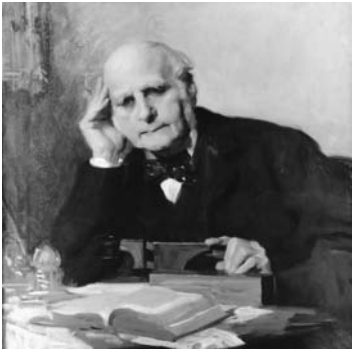
This model tells us that for every standard deviation above (or below) the mean a sandwich is in protein, we'll predict that its fat content is 0.83 standard deviations above (or below) the mean fat content. A double hamburger has 31 grams of protein, about 1 SD above the mean. Putting 1.0 in for  $z_{Protein}$  in the model gives a  $\hat{z}_{Fat}$  value of 0.83. If you trust the model, you'd expect the fat content to be about 0.83 fat SDs above the mean fat level. **Moving one standard deviation away from the mean in  $x$  moves our estimate  $r$  standard deviations away from the mean in  $y$ .**

If  $r = 0$ , there's no linear relationship. The line is horizontal, and no matter how many standard deviations you move in  $x$ , the predicted value for  $y$  doesn't



change. On the other hand, if  $r = 1.0$  or  $-1.0$ , there's a perfect linear association. In that case, moving any number of standard deviations in  $x$  moves exactly the same number of standard deviations in  $y$ . In general, moving any number of standard deviations in  $x$  moves  $r$  times that number of standard deviations in  $y$ .

## How Big Can Predicted Values Get?



Sir Francis Galton was the first to speak of “regression,” although others had fit lines to data by the same method.

### The First Regression

Sir Francis Galton related the heights of sons to the heights of their fathers with a regression line. The slope of his line was less than 1. That is, sons of tall fathers were tall, but not as much above the average height as their fathers had been above their mean. Sons of short fathers were short, but generally not as far from their mean as their fathers. Galton interpreted the slope correctly as indicating a “regression” toward the mean height—and “regression” stuck as a description of the method he had used to find the line.

Suppose you were told that a new male student was about to join the class, and you were asked to guess his height in inches. What would be your guess? A safe guess would be the mean height of male students. Now suppose you are also told that this student has a grade point average (*GPA*) of 3.9—about 2 SDs above the mean *GPA*. Would that change your guess? Probably not. The correlation between *GPA* and *height* is near 0, so knowing the *GPA* value doesn't tell you anything and doesn't move your guess. (And the equation tells us that as well, since it says that we should move  $0 \times 2$  SDs from the mean.)

On the other hand, suppose you were told that, measured in centimeters, the student's height was 2 SDs above the mean. There's a perfect correlation between *height in inches* and *height in centimeters*, so you'd know he's 2 SDs above mean height in inches as well. (The equation would tell us to move  $1.0 \times 2$  SDs from the mean.)

What if you're told that the student is 2 SDs above the mean in *shoe size*? Would you still guess that he's of average *height*? You might guess that he's taller than average, since there's a positive correlation between *height* and *shoe size*. But would you guess that he's 2 SDs above the mean? When there was no correlation, we didn't move away from the mean at all. With a perfect correlation, we moved our guess the full 2 SDs. Any correlation between these extremes should lead us to move somewhere between 0 and 2 SDs above the mean. (To be exact, the equation tells us to move  $r \times 2$  standard deviations away from the mean.)

Notice that if  $x$  is 2 SDs above its mean, we won't ever guess more than 2 SDs away for  $y$ , since  $r$  can't be bigger than 1.0.<sup>2</sup> So, each predicted  $y$  tends to be closer to its mean (in standard deviations) than its corresponding  $x$  was. This property of the linear model is called **regression to the mean**, and the line is called the **regression line**.



### JUST CHECKING

A scatterplot of house *Price* (in thousands of dollars) vs. house *Size* (in thousands of square feet) for houses sold recently in Saratoga, NY shows a relationship that is straight, with only moderate scatter and no outliers. The correlation between house *Price* and house *Size* is 0.77.

1. You go to an open house and find that the house is 1 standard deviation above the mean in size. What would you guess about its price?
2. You read an ad for a house priced 2 standard deviations below the mean. What would you guess about its size?
3. A friend tells you about a house whose size in square meters (he's European) is 1.5 standard deviations above the mean. What would you guess about its size in square feet?

<sup>2</sup> In the last chapter we asserted that correlations max out at 1, but we never actually *proved* that. Here's yet another reason to check out the Math Box on the next page.

## MATH BOX

Where does the equation of the line of best fit come from? To write the equation of any line, we need to know a point on the line and the slope. The point is easy. Consider the BK menu example. Since it is logical to predict that a sandwich with average protein will contain average fat, the line passes through the point  $(\bar{x}, \bar{y})$ .<sup>3</sup>

To think about the slope, we look once again at the  $z$ -scores. We need to remember a few things:

1. The mean of any set of  $z$ -scores is 0. This tells us that the line that best fits the  $z$ -scores passes through the origin  $(0,0)$ .

2. The standard deviation of a set of  $z$ -scores is 1, so the variance is also 1. This means that

$$\frac{\sum (z_y - \bar{z}_y)^2}{n - 1} = \frac{\sum (z_y - 0)^2}{n - 1} = \frac{\sum z_y^2}{n - 1} = 1, \text{ a fact that will be important soon.}$$

3. The correlation is  $r = \frac{\sum z_x z_y}{n - 1}$ , also important soon.

Ready? Remember that our objective is to find the slope of the best fit line. Because it passes through the origin, its equation will be of the form  $\hat{z}_y = mz_x$ . We want to find the value for  $m$  that will minimize the sum of the squared residuals. Actually we'll divide that sum by  $n - 1$  and minimize this "mean squared residual," or *MSR*. Here goes:

$$\text{Minimize:} \quad MSR = \frac{\sum (z_y - \hat{z}_y)^2}{n - 1}$$

$$\text{Since } \hat{z}_y = mz_x: \quad MSR = \frac{\sum (z_y - mz_x)^2}{n - 1}$$

$$\text{Square the binomial:} \quad = \frac{\sum (z_y^2 - 2mz_x z_y + m^2 z_x^2)}{n - 1}$$

$$\text{Rewrite the summation:} \quad = \frac{\sum z_y^2}{n - 1} - 2m \frac{\sum z_x z_y}{n - 1} + m^2 \frac{\sum z_x^2}{n - 1}$$

$$4. \text{ Substitute from (2) and (3):} \quad = 1 - 2mr + m^2$$

Wow! That simplified nicely! And as a bonus, the last expression is quadratic. Remember parabolas from algebra class? A parabola in the form  $y = ax^2 + bx + c$  reaches its minimum at

its turning point, which occurs when  $x = \frac{-b}{2a}$ . We can minimize the mean of squared residuals

$$\text{by choosing } m = \frac{-(-2r)}{2(1)} = r.$$

Wow, again! The slope of the best fit line for  $z$ -scores is the correlation,  $r$ . This stunning fact immediately leads us to two important additional results, listed below. As you read on in the text, we explain them in the context of our continuing discussion of Burger King foods.

- A slope of  $r$  for  $z$ -scores means that for every increase of 1 standard deviation in  $z_x$  there is an increase of  $r$  standard deviations in  $\hat{z}_y$ . "Over one, up  $r$ ," as you probably said in algebra class. Translate that back to the original  $x$  and  $y$  values: "Over one standard deviation in  $x$ , up  $r$  standard deviations in  $\hat{y}$ ."

$$\text{That's it! In } x\text{- and } y\text{-values, the slope of the regression line is } b = \frac{rs_y}{s_x}.$$

<sup>3</sup> It's actually not hard to prove this too.

- We know choosing  $m = r$  minimizes the sum of the squared residuals, but how small does that sum get? Equation (4) told us that the mean of the squared residuals is  $1 - 2mr + m^2$ . When  $m = r$ ,  $1 - 2mr + m^2 = 1 - 2r^2 + r^2 = 1 - r^2$ . This is the variability *not* explained by the regression line. Since the variance in  $z_y$  was 1 (Equation 2), the percentage of variability in  $y$  that is explained by  $x$  is  $r^2$ . This important fact will help us assess the strength of our models.

And there's still another bonus. Because  $r^2$  is the percent of variability explained by our model,  $r^2$  is at most 100%. If  $r^2 \leq 1$ , then  $-1 \leq r \leq 1$ , proving that correlations are always between  $-1$  and  $+1$ . (Told you so!)

## The Regression Line in Real Units

### Why Is Correlation “ $r$ ”?

In his original paper on correlation, Galton used  $r$  for the “index of correlation” that we now call the correlation coefficient. He calculated it from the regression of  $y$  on  $x$  or of  $x$  on  $y$  after standardizing the variables, just as we have done. It's fairly clear from the text that he used  $r$  to stand for (standardized) regression.

**AS** **Simulation: Interpreting Equations.** This demonstrates how to use and interpret linear equations.

Protein	Fat
$\bar{x} = 17.2$ g	$\bar{y} = 23.5$ g
$s_x = 14.0$ g	$s_y = 16.4$ g
$r = 0.83$	

### Slope

$$b_1 = \frac{rs_y}{s_x}$$

### Intercept

$$b_0 = \bar{y} - b_1\bar{x}$$

When you read the Burger King menu, you probably don't think in  $z$ -scores. But you might want to know the fat content in grams for a specific amount of protein in grams.

How much fat should we predict for a double hamburger with 31 grams of protein? The mean protein content is near 17 grams and the standard deviation is 14, so that item is 1 SD above the mean. Since  $r = 0.83$ , we predict the fat content will be 0.83 SDs above the mean fat content. Great. How much fat is that? Well, the mean fat content is 23.5 grams and the standard deviation of fat content is 16.4, so we predict that the double hamburger will have  $23.5 + 0.83 \times 16.4 = 37.11$  grams of fat.

We can always convert both  $x$  and  $y$  to  $z$ -scores, find the correlation, use  $\hat{z}_y = rz_x$  and then convert  $\hat{z}_y$  back to its original units so that we can understand the prediction. But can't we do this more simply?

Yes. Let's write the equation of the line for protein and fat—that is, the actual  $x$  and  $y$  values rather than their  $z$ -scores. In Algebra class you may have once seen lines written in the form  $y = mx + b$ . Statisticians do exactly the same thing, but with different notation:

$$\hat{y} = b_0 + b_1x.$$

In this equation,  $b_0$  is the  **$y$ -intercept**, the value of  $y$  where the line crosses the  $y$ -axis, and  $b_1$  is the **slope**.<sup>4</sup>

First we find the slope, using the formula we developed in the Math Box.<sup>5</sup> Remember? We know that our model predicts that for each increase of one standard deviation in protein we'll see an increase of about 0.83 standard deviations in fat.

In other words, the slope of the line in original units is

$$b_1 = \frac{rs_y}{s_x} = \frac{0.83 \times 16.4 \text{ g fat}}{14 \text{ g protein}} = 0.97 \text{ grams of fat per gram of protein.}$$

Next, how do we find the  $y$ -intercept,  $b_0$ ? Remember that the line has to go through the mean-mean point  $(\bar{x}, \bar{y})$ . In other words, the model predicts  $\bar{y}$  to be the value that corresponds to  $\bar{x}$ . We can put the means into the equation and write  $\bar{y} = b_0 + b_1\bar{x}$ .

Solving for  $b_0$ , we see that the intercept is just  $b_0 = \bar{y} - b_1\bar{x}$ .

<sup>4</sup> We changed from  $mx + b$  to  $b_0 + b_1x$  for a reason—not just to be difficult. Eventually we'll want to add more  $x$ 's to the model to make it more realistic and we don't want to use up the entire alphabet. What would we use after  $m$ ? The next letter is  $n$ , and that one's already taken.  $o$ ? See our point? Sometimes subscripts are the best approach.

<sup>5</sup> Several important results popped up in that Math Box. Check it out!

For the Burger King foods, that comes out to

$$b_0 = 23.5 \text{ g fat} - 0.97 \frac{\text{g fat}}{\text{g protein}} \times 17.2 \text{ g protein} = 6.8 \text{ g fat.}$$

Putting this back into the regression equation gives

$$\widehat{\text{fat}} = 6.8 + 0.97 \text{ protein.}$$

### Units of $y$ per unit of $x$

Get into the habit of identifying the units by writing down “ $y$ -units per  $x$ -unit,” with the unit names put in place. You’ll find it’ll really help you to tell about the line in context.

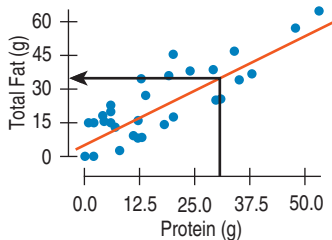


FIGURE 8.4

Burger King menu items in their natural units with the regression line.

What does this mean? The slope, 0.97, says that an additional gram of protein is associated with an additional 0.97 grams of fat, on average. Less formally, we might say that Burger King sandwiches pack about 0.97 grams of fat per gram of protein. Slopes are always expressed in  $y$ -units per  $x$ -unit. They tell how the  $y$ -variable changes (in its units) for a one-unit change in the  $x$ -variable. When you see a phrase like “students per teacher” or “kilobytes per second” think slope.

Changing the units of the variables doesn’t change the *correlation*, but for the *slope*, units do matter. We may know that age and height in children are positively correlated, but the *value* of the slope depends on the units. If children grow an average of 3 inches per year, that’s the same as 0.21 millimeters per day. For the slope, it matters whether you express age in days or years and whether you measure height in inches or millimeters. How you choose to express  $x$  and  $y$ —what units you use—affects the slope directly. Why? We know changing units doesn’t change the correlation, but does change the standard deviations. The slope introduces the units into the equation by multiplying the correlation by the ratio of  $s_y$  to  $s_x$ . **The units of the slope are always the units of  $y$  per unit of  $x$ .**

How about the **intercept** of the BK regression line, 6.8? Algebraically, that’s the value the line takes when  $x$  is zero. Here, our model predicts that even a BK item with no protein would have, on average, about 6.8 grams of fat. Is that reasonable? Well, the apple pie, with 2 grams of protein, has 14 grams of fat, so it’s not impossible. But often 0 is not a plausible value for  $x$  (the year 0, a baby born weighing 0 grams, ...). Then the intercept serves only as a starting value for our predictions and we don’t interpret it as a meaningful predicted value.

## FOR EXAMPLE

### A regression model for hurricanes

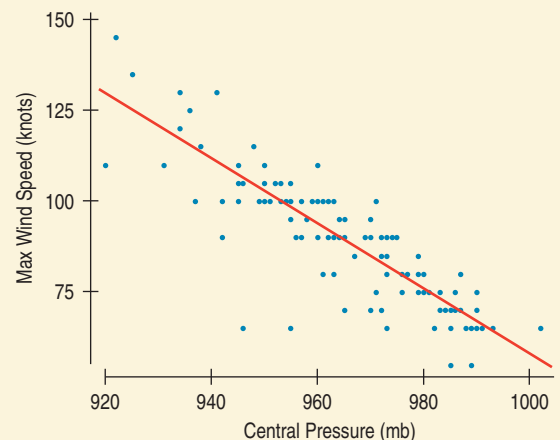
In Chapter 7 we looked at the relationship between the central pressure and maximum wind speed of Atlantic hurricanes. We saw that the scatterplot was straight enough, and then found a correlation of  $-0.879$ , but we had no model to describe how these two important variables are related or to allow us to predict wind speed from pressure. Since the conditions we need to check for regression are the same ones we checked before, we can use technology to find the regression model. It looks like this:

$$\widehat{\text{MaxWindSpeed}} = 955.27 - 0.897 \text{ CentralPressure}$$

**Question:** Interpret this model. What does the slope mean in this context? Does the intercept have a meaningful interpretation?

The negative slope says that as *CentralPressure* falls, *MaxWindSpeed* increases. That makes sense from our general understanding of how hurricanes work: Low central pressure pulls in moist air, driving the rotation and the resulting destructive winds. The slope’s value says that, on average, the maximum wind speed increases by about 0.897 knots for every 1-millibar drop in central pressure.

It’s not meaningful, however, to interpret the intercept as the wind speed predicted for a central pressure of 0—that would be a vacuum. Instead, it is merely a starting value for the model.



With the estimated linear model,  $\widehat{fat} = 6.8 + 0.97 \text{ protein}$ , it's easy to predict fat content for any menu item we want. For example, for the BK Broiler chicken sandwich with 30 grams of *protein*, we can plug in 30 grams for the amount of *protein* and see that the *predicted fat* content is  $6.8 + 0.97(30) = 35.9$  grams of fat. Because the BK Broiler chicken sandwich actually has 25 grams of fat, its residual is

$$fat - \widehat{fat} = 25 - 35.9 = -10.9 \text{ g.}$$

To use a regression model, we should check the same conditions for regressions as we did for correlation: the **Quantitative Variables Condition**, the **Straight Enough Condition**, and the **Outlier Condition**.



### JUST CHECKING

Let's look again at the relationship between house *Price* (in thousands of dollars) and house *Size* (in thousands of square feet) in Saratoga. The regression model is

$$\widehat{Price} = -3.117 + 94.454 \text{ Size.}$$

4. What does the slope of 94.454 mean?
5. What are the units of the slope?
6. Your house is 2000 sq ft bigger than your neighbor's house. How much more do you expect it to be worth?
7. Is the  $y$ -intercept of  $-3.117$  meaningful? Explain.

### STEP-BY-STEP EXAMPLE

### Calculating a Regression Equation



Wildfires are an ongoing source of concern shared by several government agencies. In 2004, the Bureau of Land Management, Bureau of Indian Affairs, Fish and Wildlife Service, National Park Service, and USDA Forest Service spent a combined total of \$890,233,000 on fire suppression, down from nearly twice that much in 2002. These government agencies join together in the National Interagency Fire Center, whose Web site ([www.nifc.gov](http://www.nifc.gov)) reports statistics about wildfires.

**Question:** Has the annual number of wildfires been changing, on average? If so, how fast and in what way?



**Plan** State the problem.

**Variables** Identify the variables and report the  $W$ 's.

Check the appropriate assumptions and conditions.

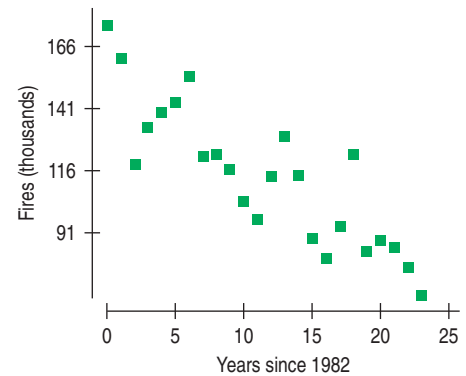
I want to know how the number of wildfires in the continental United States has changed in the past two decades.

I have data giving the number of wildfires for each year (in thousands of fires) from 1982 to 2005.

✓ **Quantitative Variables Condition:** Both the number of fires and the year are quantitative.

Just as we did for correlation, check the conditions for a regression by making a picture. Never fit a regression without looking at the scatterplot first.

*Note:* It's common (and usually simpler) not to use four-digit numbers to identify years. Here we have chosen to number the years beginning in 1982, so 1982 is represented as year 0 and 2005 as year 23.



- ✓ **Straight Enough Condition:** The scatterplot shows a strong linear relationship with a negative association.
- ✓ **Outlier Condition:** No outliers are evident in the scatterplot.

Because these conditions are satisfied, it is OK to model the relationship with a regression line.

SHOW

**Mechanics** Find the equation of the regression line. Summary statistics give the building blocks of the calculation.

(We generally report summary statistics to one more digit of accuracy than the data. We do the same for intercept and predicted values, but for slopes we usually report an additional digit. Remember, though, not to round off until you finish computing an answer.)<sup>6</sup>

Find the slope,  $b_1$ .

Find the intercept,  $b_0$ .

Write the equation of the model, using meaningful variable names.

**Year:**

$$\bar{x} = 11.5 \text{ (representing 1993.5)}$$

$$s_x = 7.07 \text{ years}$$

**Fires:**

$$\bar{y} = 114.098 \text{ fires}$$

$$s_y = 28.342 \text{ fires}$$

**Correlation:**

$$r = -0.862$$

$$b_1 = \frac{rs_y}{s_x} = \frac{-0.862(28.342)}{7.07}$$

$$= -3.4556 \text{ fires per year}$$

$$b_0 = y - b_1x = 114.098 - (-3.4556)11.5$$

$$= 153.837$$

So the least squares line is

$$\hat{y} = 153.837 - 3.4556x, \text{ or}$$

$$\widehat{\text{Fires}} = 153.837 - 3.4556 \text{ year}$$

<sup>6</sup> We warned you in Chapter 6 that we'll round in the intermediate steps of a calculation to show the steps more clearly. If you repeat these calculations yourself on a calculator or statistics program, you may get somewhat different results. When calculated with more precision, the intercept is 153,809 and the slope is  $-3.453$ .



**Conclusion** Interpret what you have found in the context of the question. Discuss in terms of the variables and their units.

During the period from 1982 to 2005, the annual number of fires declined at an average rate of about 3,456 (3.456 thousand) fires per year. For prediction, the model uses a base estimation of 153,837 fires in 1982.

**AS** **Activity:** Find a **Regression Equation.** Now that we've done it by hand, try it with technology using the statistics package paired with your version of *ActivStats*.

## Residuals Revisited

### Why $e$ for "Residual"?

The flip answer is that  $r$  is already taken, but the truth is that  $e$  stands for "error." No, that doesn't mean it's a mistake. Statisticians often refer to variability not explained by a model as error.

The linear model we are using assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled. We can write

$$\text{Data} = \text{Model} + \text{Residual}$$

or, equivalently,

$$\text{Residual} = \text{Data} - \text{Model}.$$

Or, in symbols,

$$e = y - \hat{y}.$$

When we want to know how well the model fits, we can ask instead what the model missed. To see that, we look at the residuals.

### FOR EXAMPLE

#### Katrina's residual

**Recap:** The linear model relating hurricanes' wind speeds to their central pressures was

$$\widehat{\text{MaxWindSpeed}} = 955.27 - 0.897\text{CentralPressure}$$

Let's use this model to make predictions and see how those predictions do.

**Question:** Hurricane Katrina had a central pressure measured at 920 millibars. What does our regression model predict for her maximum wind speed? How good is that prediction, given that Katrina's actual wind speed was measured at 110 knots?

Substituting 920 for the central pressure in the regression model equation gives

$$\widehat{\text{MaxWindSpeed}} = 955.27 - 0.897(920) = 130.03$$

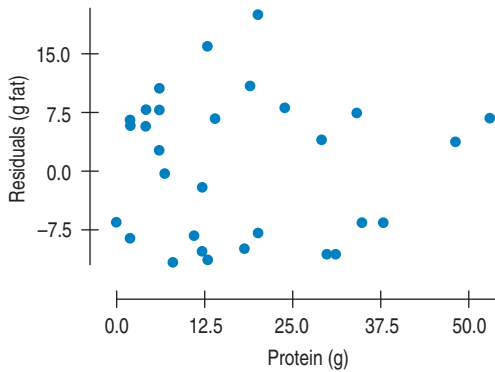
The regression model predicts a maximum wind speed of 130 knots for Hurricane Katrina.

The residual for this prediction is the observed value minus the predicted value:

$$110 - 130 = -20\text{kts}.$$

In the case of Hurricane Katrina, the model predicts a wind speed 20 knots higher than was actually observed.





**FIGURE 8.5**

The residuals for the BK menu regression look appropriately boring.

Residuals help us to see whether the model makes sense. When a regression model is appropriate, it should model the underlying relationship. Nothing interesting should be left behind. So after we fit a regression model, we usually plot the residuals in the hope of finding . . . nothing.

A scatterplot of the residuals versus the  $x$ -values should be the most boring scatterplot you've ever seen. It shouldn't have any interesting features, like a direction or shape. It should stretch horizontally, with about the same amount of scatter throughout. It should show no bends, and it should have no outliers. If you see any of these features, find out what the regression model missed.

Most computer statistics packages plot the residuals against the predicted values  $\hat{y}$ , rather than against  $x$ . When the slope is negative, the two versions are mirror images. When the slope is positive, they're virtually identical except for the axis labels. Since all we care about is the patterns (or, better, lack of patterns) in the plot, it really doesn't matter which way we plot the residuals.



### JUST CHECKING

Our linear model for Saratoga homes uses the *Size* (in thousands of square feet) to estimate the *Price* (in thousands of dollars):  $\widehat{Price} = -3.117 + 94.454Size$ . Suppose you're thinking of buying a home there.

8. Would you prefer to find a home with a negative or a positive residual? Explain.
9. You plan to look for a home of about 3000 square feet. How much should you expect to have to pay?
10. You find a nice home that size selling for \$300,000. What's the residual?

## The Residual Standard Deviation

If the residuals show no interesting pattern when we plot them against  $x$ , we can look at how big they are. After all, we're trying to make them as small as possible. Since their mean is always zero, though, it's only sensible to look at how much they vary. The standard deviation of the residuals,  $s_e$ , gives us a measure of how much the points spread around the regression line. Of course, for this summary to make sense, the residuals should all share the same underlying spread, so we check to make sure that the residual plot has about the same amount of scatter throughout.

This gives us a new assumption: the **Equal Variance Assumption**. The associated condition to check is the **Does the Plot Thicken? Condition**. We check to make sure that the spread is about the same all along the line. We can check that either in the original scatterplot of  $y$  against  $x$  or in the scatterplot of residuals.

We estimate the standard deviation of the residuals in almost the way you'd expect:

$$s_e = \sqrt{\frac{\sum e^2}{n - 2}}$$

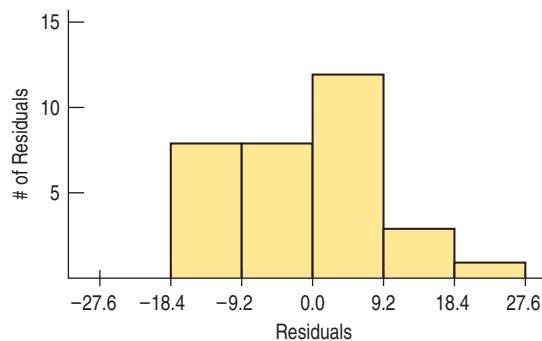
We don't need to subtract the mean because the mean of the residuals  $\bar{e} = 0$ .

For the Burger King foods, the standard deviation of the residuals is 9.2 grams of fat. That looks about right in the scatterplot of residuals. The residual for the BK Broiler chicken was  $-11$  grams, just over one standard deviation.

Why  $n - 2$  rather than  $n - 1$ ? We used  $n - 1$  for  $s$  when we estimated the mean. Now we're estimating both a slope and an intercept. Looks like a pattern—and it is. We subtract one more for each parameter we estimate.

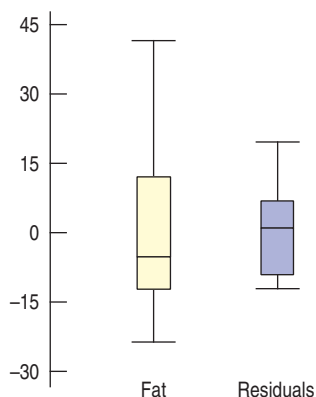


It's a good idea to make a histogram of the residuals. If we see a unimodal, symmetric histogram, then we can apply the 68–95–99.7 Rule to see how well the regression model describes the data. In particular, we know that 95% of the residuals should be no larger in size than  $2s_e$ . The Burger King residuals look like this:



Sure enough, almost all are less than  $2(9.2)$ , or 18.4, g of fat in size.

## $R^2$ —The Variation Accounted For



**FIGURE 8.6**

Compare the variability of total Fat with the residuals from the regression. The means have been subtracted to make it easier to compare spreads. The variation left in the residuals is unaccounted for by the model, but it's less than the variation in the original data.

### TI-*n*spire

**Understanding  $R^2$ .** Watch the unexplained variability decrease as you drag points closer to the regression line.

The variation in the residuals is the key to assessing how well the model fits. Let's compare the variation of the response variable with the variation of the residuals. The total *Fat* has a standard deviation of 16.4 grams. The standard deviation of the residuals is 9.2 grams. If the correlation were 1.0 and the model predicted the *Fat* values perfectly, the residuals would all be zero and have no variation. We couldn't possibly do any better than that.

On the other hand, if the correlation were zero, the model would simply predict 23.5 grams of *Fat* (the mean) for all menu items. The residuals from that prediction would just be the observed *Fat* values minus their mean. These residuals would have the same variability as the original data because, as we know, just subtracting the mean doesn't change the spread.

How well does the BK regression model do? Look at the boxplots. The variation in the residuals is smaller than in the data, but certainly bigger than zero. That's nice to know, but how much of the variation is still left in the residuals? If you had to put a number between 0% and 100% on the fraction of the variation left in the residuals, what would you say?

All regression models fall somewhere between the two extremes of zero correlation and perfect correlation. We'd like to gauge where our model falls. As we showed in the Math Box,<sup>7</sup> the squared correlation,  $r^2$ , gives the fraction of the data's variation accounted for by the model, and  $1 - r^2$  is the fraction of the original variation left in the residuals. For the Burger King model,  $r^2 = 0.83^2 = 0.69$ , and  $1 - r^2$  is 0.31, so 31% of the variability in total *Fat* has been left in the residuals. How close was that to your guess?

All regression analyses include this statistic, although by tradition, it is written with a capital letter,  $R^2$ , and pronounced "R-squared." An  $R^2$  of 0 means that none of the variance in the data is in the model; all of it is still in the residuals. It would be hard to imagine using that model for anything.

<sup>7</sup> Have you looked yet? Please do.

Is a correlation of 0.80 twice as strong as a correlation of 0.40? Not if you think in terms of  $R^2$ . A correlation of 0.80 means an  $R^2$  of  $0.80^2 = 64\%$ . A correlation of 0.40 means an  $R^2$  of  $0.40^2 = 16\%$ —only a quarter as much of the variability accounted for. A correlation of 0.80 gives an  $R^2$  four times as strong as a correlation of 0.40 and accounts for four times as much of the variability.

Because  $R^2$  is a fraction of a whole, it is often given as a percentage.<sup>8</sup> For the Burger King data,  $R^2$  is 69%. When interpreting a regression model, you need to *Tell* what  $R^2$  means. According to our linear model, 69% of the variability in the fat content of Burger King sandwiches is accounted for by variation in the protein content.

**How can we see that  $R^2$  is really the fraction of variance accounted for by the model?** It's a simple calculation. The variance of the fat content of the Burger King foods is  $16.4^2 = 268.42$ . If we treat the residuals as data, the variance of the residuals is 83.195.<sup>9</sup> As a fraction, that's  $83.195/268.42 = 0.31$ , or 31%. That's the fraction of the variance that is not accounted for by the model. The fraction that is accounted for is  $100\% - 31\% = 69\%$ , just the value we got for  $R^2$ .

## FOR EXAMPLE

### Interpreting $R^2$

**Recap:** Our regression model that predicts maximum wind speed in hurricanes based on the storm's central pressure has  $R^2 = 77.3\%$ .

**Question:** What does that say about our regression model?

An  $R^2$  of 77.3% indicates that 77.3% of the variation in maximum wind speed can be accounted for by the hurricane's central pressure. Other factors, such as temperature and whether the storm is over water or land, may explain some of the remaining variation.



## JUST CHECKING

Back to our regression of house *Price* (in thousands of \$) on house *Size* (in thousands of square feet). The  $R^2$  value is reported as 59.5%, and the standard deviation of the residuals is 53.79.

11. What does the  $R^2$  value mean about the relationship of *Price* and *Size*?
12. Is the correlation of *Price* and *Size* positive or negative? How do you know?
13. If we measure house *Size* in square meters instead, would  $R^2$  change? Would the slope of the line change? Explain.
14. You find that your house in Saratoga is worth \$100,000 more than the regression model predicts. Should you be very surprised (as well as pleased)?

## How Big Should $R^2$ Be?

$R^2$  is always between 0% and 100%. But what's a "good"  $R^2$  value? The answer depends on the kind of data you are analyzing and on what you want to do with it. Just as with correlation, there is no value for  $R^2$  that automatically determines

<sup>8</sup> By contrast, we usually give correlation coefficients as decimal values between  $-1.0$  and  $1.0$ .

<sup>9</sup> This isn't quite the same as squaring the  $s_e$  that we discussed on the previous page, but it's very close. We'll deal with the distinction in Chapter 27.

**Some Extreme Tales**

One major company developed a method to differentiate between proteins. To do so, they had to distinguish between regressions with  $R^2$  of 99.99% and 99.98%. For this application, 99.98% was not high enough.

The president of a financial services company reports that although his regressions give  $R^2$  below 2%, they are highly successful because those used by his competition are even lower.

that the regression is “good.” Data from scientific experiments often have  $R^2$  in the 80% to 90% range and even higher. Data from observational studies and surveys, though, often show relatively weak associations because it’s so difficult to measure responses reliably. An  $R^2$  of 50% to 30% or even lower might be taken as evidence of a useful regression. The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.

As we’ve seen, an  $R^2$  of 100% is a perfect fit, with no scatter around the line. The  $s_e$  would be zero. All of the variance is accounted for by the model and none is left in the residuals at all. This sounds great, but it’s too good to be true for real data.<sup>10</sup>

Along with the slope and intercept for a regression, you should always report  $R^2$  so that readers can judge for themselves how successful the regression is at fitting the data. Statistics is about variation, and  $R^2$  measures the success of the regression model in terms of the fraction of the variation of  $y$  accounted for by the regression.  $R^2$  is the first part of a regression that many people look at because, along with the scatterplot, it tells whether the regression model is even worth thinking about.

## Regression Assumptions and Conditions

The linear regression model is perhaps the most widely used model in all of Statistics. It has everything we could want in a model: two easily estimated parameters, a meaningful measure of how well the model fits the data, and the ability to predict new values. It even provides a self-check in plots of the residuals to help us avoid silly mistakes.

Like all models, though, linear models don’t apply all the time, so we’d better think about whether they’re reasonable. It makes no sense to make a scatterplot of categorical variables, and even less to perform a regression on them. Always check the **Quantitative Variables Condition** to be sure a regression is appropriate.

The linear model makes several assumptions. First, and foremost, is the **Linearity Assumption**—that the relationship between the variables is, in fact, linear. You can’t verify an assumption, but you can check the associated condition. A quick look at the scatterplot will help you check the **Straight Enough Condition**. You don’t need a *perfectly* straight plot, but it must be straight enough for the linear model to make sense. If you try to model a curved relationship with a straight line, you’ll usually get exactly what you deserve.

If the scatterplot is not straight enough, stop here. You can’t use a linear model for *any* two variables, even if they are related. They must have a *linear* association, or the model won’t mean a thing.

For the standard deviation of the residuals to summarize the scatter, all the residuals should share the same spread, so we need the **Equal Variance Assumption**. The **Does the Plot Thicken? Condition** checks for changing spread in the scatterplot.

Check the **Outlier Condition**. Outlying points can dramatically change a regression model. Outliers can even change the sign of the slope, misleading us about the underlying relationship between the variables. We’ll see examples in the next chapter.

Even though we’ve checked the conditions in the scatterplot of the data, a scatterplot of the residuals can sometimes help us see any violations even more

**Make a Picture**

To use regression, first check that

- the scatterplot is straight enough.

After you’ve fit the regression, make a residual plot and check that there are no obvious patterns. In particular, check that

- there are no obvious bends,
- the spread of the residuals is about the same throughout, and
- there are no obvious outliers.

<sup>10</sup> If you see an  $R^2$  of 100%, it’s a good idea to figure out what happened. You may have discovered a new law of Physics, but it’s much more likely that you accidentally regressed two variables that measure the same thing.

clearly. And examining the residuals is the best way to look for additional patterns and interesting quirks in the data.

## A Tale of Two Regressions

Regression slopes may not behave exactly the way you'd expect at first. Our regression model for the Burger King sandwiches was  $\widehat{fat} = 6.8 + 0.97 \text{ protein}$ . That equation allowed us to estimate that a sandwich with 30 grams of protein would have 35.9 grams of fat. Suppose, though, that we knew the fat content and wanted to predict the amount of protein. It might seem natural to think that by solving our equation for *protein* we'd get a model for predicting *protein* from *fat*. But that doesn't work.

Our original model is  $\hat{y} = b_0 + b_1x$ , but the new one needs to evaluate an  $\hat{x}$  based on a value of  $y$ . There's no  $y$  in our original model, only  $\hat{y}$ , and that makes all the difference. Our model doesn't fit the BK data values perfectly, and the least squares criterion focuses on the vertical errors the model makes in using to model  $y$ —not on horizontal errors related to  $x$ .

A quick look at the equations reveals why. Simply solving our equation for  $x$  would give a new line whose slope is the reciprocal of ours. To model  $y$  in terms of  $x$ , our slope is  $b_1 = \frac{rs_y}{s_x}$ . To model  $x$  in terms of  $y$ , we'd need to use the slope  $b_1 = \frac{rs_x}{s_y}$ . Notice that it's *not* the reciprocal of ours.

If we want to predict *protein* from *fat*, we need to create that model. The slope is  $b_1 = \frac{(0.83)(14.0)}{16.4} = 0.709$  grams of protein per gram of fat. The equation turns out to be  $\widehat{protein} = 0.55 + 0.709 \text{ fat}$ , so we'd predict that a sandwich with 35.9 grams of fat should have 26.0 grams of protein—not the 30 grams that we used in the first equation.

Moral of the story: *Think*. (Where have you heard *that* before?) Decide which variable you want to use ( $x$ ) to predict values for the other ( $y$ ). Then find the model that does that. If, later, you want to make predictions in the other direction, you'll need to start over and create the other model from scratch.

Protein	Fat
$\bar{x} = 17.2 \text{ g}$	$\bar{y} = 23.5 \text{ g}$
$s_x = 14.0 \text{ g}$	$s_y = 16.4 \text{ g}$
$r = 0.83$	

### STEP-BY-STEP EXAMPLE

### Regression

Even if you hit the fast food joints for lunch, you should have a good breakfast. Nutritionists, concerned about “empty calories” in breakfast cereals, recorded facts about 77 cereals, including their *Calories* per serving and *Sugar* content (in grams).

**Question:** How are calories and sugar content related in breakfast cereals?



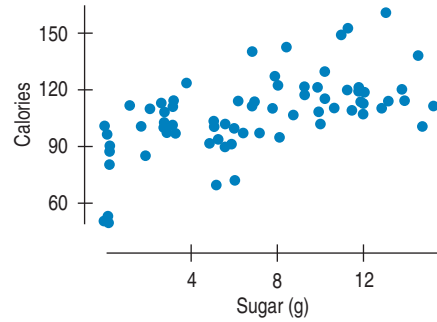
**Plan** State the problem and determine the role of the variables.

**Variables** Name the variables and report the  $W$ 's.

I am interested in the relationship between sugar content and calories in cereals. I'll use *Sugar* to estimate *Calories*.

✓ **Quantitative Variables Condition:** I have two quantitative variables, *Calories* and *Sugar* content per serving, measured on 77 breakfast cereals. The units of measurement are calories and grams of sugar, respectively.

Check the conditions for a regression by making a picture. Never fit a regression without looking at the scatterplot first.



- ✓ **Outlier Condition:** There are no obvious outliers or groups.
- ✓ The **Straight Enough Condition** is satisfied; I will fit a regression model to these data.
- ✓ The **Does the Plot Thicken? Condition** is satisfied. The spread around the line looks about the same throughout.

SHOW

**Mechanics** If there are no clear violations of the conditions, fit a straight line model of the form  $\hat{y} = b_0 + b_1x$  to the data. Summary statistics give the building blocks of the calculation.

Find the slope.

Find the intercept.

Write the equation, using meaningful variable names.

State the value of  $R^2$ .

**Calories**

$$\bar{y} = 107.0 \text{ calories}$$

$$s_y = 19.5 \text{ calories}$$

**Sugar**

$$\bar{x} = 7.0 \text{ grams}$$

$$s_x = 4.4 \text{ grams}$$

**Correlation**

$$r = 0.564$$

$$b_1 = \frac{rs_y}{s_x} = \frac{0.564(19.5)}{4.4}$$

$$= 2.50 \text{ calories per gram of sugar.}$$

$$b_0 = \bar{y} - b_1\bar{x} = 107 - 2.50(7) = 89.5 \text{ calories.}$$

So the least squares line is

$$\widehat{\text{Calories}} = 89.5 + 2.50 \text{ Sugar.}$$

Squaring the correlation gives

$$R^2 = 0.564^2 = 0.318 \text{ or } 31.8\%.$$

TELL

**Conclusion** Describe what the model says in words and numbers. Be sure to use the names of the variables and their units.

The key to interpreting a regression model is to start with the phrase “ $b_1$   $y$ -units per  $x$ -unit,” substituting the estimated value of the slope for  $b_1$  and the names of the

The scatterplot shows a positive, linear relationship and no outliers. The slope of the least squares regression line suggests that cereals have about 2.50 Calories more per additional gram of Sugar.

respective units. The intercept is then a starting or base value.

$R^2$  gives the fraction of the variability of  $y$  accounted for by the linear regression model.

Find the standard deviation of the residuals,  $s_e$ , and compare it to the original  $s_y$ .

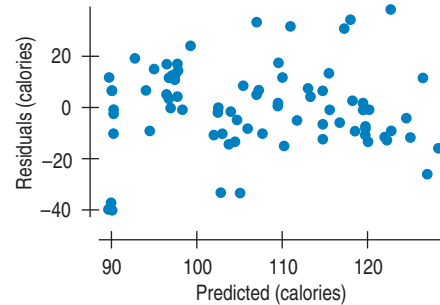
The intercept predicts that sugar-free cereals would average about 89.5 calories.

The  $R^2$  says that 31.8% of the variability in Calories is accounted for by variation in Sugar content.

$s_e = 16.2$  calories. That's smaller than the original SD of 19.5, but still fairly large.



**Check Again** Even though we looked at the scatterplot *before* fitting a regression model, a plot of the residuals is essential to any regression analysis because it is the best check for additional patterns and interesting quirks in the data.



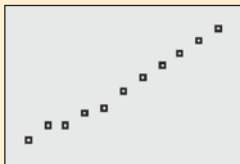
The residuals show a horizontal direction, a shapeless form, and roughly equal scatter for all predicted values. The linear model appears to be appropriate.

### TI-*mspire*

**Residuals plots.** See how the residuals plot changes as you drag points around in a scatterplot.

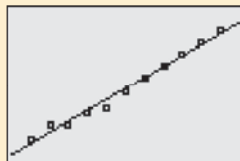
### TI Tips

### Regression lines and residuals plots



```
LinReg(a+bx) LVR
,LTUIT,Y1
```

```
LinReg
y=a+bx
a=6439.954545
b=326.0818182
r^2=.9863642357
r=.9931587163
```



By now you will not be surprised to learn that your calculator can do it all: scatterplot, regression line, and residuals plot. Let's try it using the Arizona State tuition data from the last chapter. (TI Tips, p. 149) You should still have that saved in lists named **YR** and **TUIT**. First, recreate the scatterplot.

#### 1. Find the equation of the regression line.

Actually, you already found the line when you used the calculator to get the correlation. But this time we'll be a little fancier so that we can display the line on our scatterplot. We want to tell the calculator to do the regression and save the equation of the model as a graphing variable.

- Under **STAT CALC** choose **LinReg(a+bx)**.
- Specify that  $x$  and  $y$  are **YR** and **TUIT**, as before, but . . .
- Now add a comma and one more specification. Press **VAR**, go to the **Y-VARS** menu, choose **1:Function**, and finally(!) choose **Y1**.
- Hit **ENTER**.

There's the equation. The calculator tells you that the regression line is  $\widehat{tuit} = 6440 + 326 \text{ year}$ . Can you explain what the slope and  $y$ -intercept mean?

#### 2. Add the line to the plot.

When you entered this command, the calculator automatically saved the equation as **Y1**. Just hit **GRAPH** to see the line drawn across your scatterplot.

YR	TUIT	RESID
0	6546	106.05
1	6996	229.96
2	6996	-96.12
3	7350	-68.2
4	7500	-244.3
5	7878	-82.36
6	8377	-19.45

RESID =  $\{106.04545\dots\}$

```

Plot1  [Off] Plot3
Type:  [Off] [On] [On]
Xlist: YR
Ylist: RESID
Mark:  [ ] [ ] [ ]

```

```

Plot1  [Off] Plot3
Y1=6439.9545454
545+326.08181818
182X
V2= [ ]
V3= [ ]
V4= [ ]
V5= [ ]

```



### 3. Check the residuals.

Remember, you are not finished until you check to see if a linear model is appropriate. That means you need to see if the residuals appear to be randomly distributed. To do that, you need to look at the residuals plot.

This is made easy by the fact that the calculator has already placed the residuals in a list named **RESID**. Want to see them? Go to **STAT EDIT** and look through the lists. (If **RESID** is not already there, go to the first blank list and import the name **RESID** from your **LIST NAMES** menu. The residuals should appear.) Every time you have the calculator compute a regression analysis, it will automatically save this list of residuals for you.

### 4. Now create the residuals plot.

- Set up **STAT PLOT Plot2** as a scatterplot with **Xlist:YR** and **Ylist:RESID**.
- Before you try to see the plot, go to the **V=** screen. By moving the cursor around and hitting **ENTER** in the appropriate places you can turn off the regression line and **Plot1**, and turn on **Plot2**.
- **ZoomStat** will now graph the residuals plot.

Uh-oh! See the curve? The residuals are high at both ends, low in the middle. Looks like a linear model may not be appropriate after all. Notice that the residuals plot makes the curvature much clearer than the original scatterplot did.

*Moral: Always check the residuals plot!*

So a linear model might not be appropriate here. What now? The next two chapters provide techniques for dealing with data like these.

## Reality Check: Is the Regression Reasonable?

### Adjective, Noun, or Verb

You may see the term *regression* used in different ways. There are many ways to fit a line to data, but the term “regression line” or “regression” without any other qualifiers always means least squares. People also use *regression* as a verb when they speak of *regressing* a *y*-variable on an *x*-variable to mean fitting a linear model.

Statistics don’t come out of nowhere. They are based on data. The results of a statistical analysis should reinforce your common sense, not fly in its face. If the results are surprising, then either you’ve learned something new about the world or your analysis is wrong.

Whenever you perform a regression, think about the coefficients and ask whether they make sense. Is a slope of 2.5 calories per gram of sugar reasonable? That’s hard to say right off. We know from the summary statistics that a typical cereal has about 100 calories and 7 grams of sugar. A gram of sugar contributes some calories (actually, 4, but you don’t need to know that), so calories should go up with increasing sugar. The direction of the slope seems right.

To see if the *size* of the slope is reasonable, a useful trick is to consider its order of magnitude. We’ll start by asking if deflating the slope by a factor of 10 seems reasonable. Is 0.25 calories per gram of sugar enough? The 7 grams of sugar found in the average cereal would contribute less than 2 calories. That seems too small.

Now let’s try inflating the slope by a factor of 10. Is 25 calories per gram reasonable? Then the average cereal would have 175 calories from sugar alone. The average cereal has only 100 calories per serving, though, so that slope seems too big.

We have tried inflating the slope by a factor of 10 and deflating it by 10 and found both to be unreasonable. So, like Goldilocks, we’re left with the value in the middle that’s just right. And an increase of 2.5 calories per gram of sugar is certainly *plausible*.

The small effort of asking yourself whether the regression equation is plausible is repaid whenever you catch errors or avoid saying something silly or absurd about the data. It’s too easy to take something that comes out of a computer at face value and assume that it makes sense.

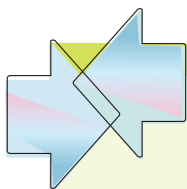
Always be skeptical and ask yourself if the answer is reasonable.

## WHAT CAN GO WRONG?

There are many ways in which data that appear at first to be good candidates for regression analysis may be unsuitable. And there are ways that people use regression that can lead them astray. Here's an overview of the most common problems. We'll discuss them at length in the next chapter.

- ▶ **Don't fit a straight line to a nonlinear relationship.** Linear regression is suited only to relationships that are, well, *linear*. Fortunately, we can often improve the linearity easily by using re-expression. We'll come back to that topic in Chapter 10.
- ▶ **Beware of extraordinary points.** Data points can be extraordinary in a regression in two ways: They can have  $y$ -values that stand off from the linear pattern suggested by the bulk of the data, or extreme  $x$ -values. Both kinds of extraordinary points require attention.
- ▶ **Don't extrapolate beyond the data.** A linear model will often do a reasonable job of summarizing a relationship in the narrow range of observed  $x$ -values. Once we have a working model for the relationship, it's tempting to use it. But beware of predicting  $y$ -values for  $x$ -values that lie outside the range of the original data. The model may no longer hold there, so such *extrapolations* too far from the data are dangerous.
- ▶ **Don't infer that  $x$  causes  $y$  just because there is a good linear model for their relationship.** When two variables are strongly correlated, it is often tempting to assume a causal relationship between them. Putting a regression line on a scatterplot tempts us even further, but it doesn't make the assumption of causation any more valid. For example, our regression model predicting hurricane wind speeds from the central pressure was reasonably successful, but the relationship is very complex. It is reasonable to say that low central pressure at the eye is responsible for the high winds because it draws moist, warm air into the center of the storm, where it swirls around, generating the winds. But as is often the case, things aren't quite that simple. The winds themselves also contribute to lowering the pressure at the center of the storm as it becomes a hurricane. Understanding causation requires far more work than just finding a correlation or modeling a relationship.
- ▶ **Don't choose a model based on  $R^2$  alone.** Although  $R^2$  measures the *strength* of the linear association, a high  $R^2$  does not demonstrate the *appropriateness* of the regression. A single outlier, or data that separate into two groups rather than a single cloud of points, can make  $R^2$  seem quite large when, in fact, the linear regression model is simply inappropriate. Conversely, a low  $R^2$  value may be due to a single outlier as well. It may be that most of the data fall roughly along a straight line, with the exception of a single point. Always look at the scatterplot.

$R^2$  does not mean that protein accounts for 69% of the fat in a BK food item. It is the *variation* in fat content that is accounted for by the linear model.



## CONNECTIONS

We've talked about the importance of models before, but have seen only the Normal model as an example. The linear model is one of the most important models in Statistics. Chapter 7 talked about the assignment of variables to the  $y$ - and  $x$ -axes. That didn't matter to correlation, but it does matter to regression because  $y$  is predicted by  $x$  in the regression model.

The connection of  $R^2$  to correlation is obvious, although it may not be immediately clear that just by squaring the correlation we can learn the fraction of the variability of  $y$  accounted for by a regression on  $x$ . We'll return to this in subsequent chapters.

We made a big fuss about knowing the units of your quantitative variables. We didn't need units for correlation, but without the units we can't define the slope of a regression. A regression makes no sense if you don't know the *Who*, the *What*, and the *Units* of both your variables.

We've summed squared deviations before when we computed the standard deviation and variance. That's not coincidental. They are closely connected to regression.

When we first talked about models, we noted that deviations away from a model were often interesting. Now we have a formal definition of these deviations as residuals.





## WHAT HAVE WE LEARNED?

We've learned that when the relationship between quantitative variables is fairly straight, a linear model can help summarize that relationship and give us insights about it:

- ▶ The regression (best fit) line doesn't pass through all the points, but it is the best compromise in the sense that the sum of squares of the residuals is the smallest possible.

We've learned several things the correlation,  $r$ , tells us about the regression:

- ▶ The slope of the line is based on the correlation, adjusted for the units of  $x$  and  $y$ :

$$b_1 = \frac{rs_y}{s_x}$$

We've learned to interpret that slope in context:

- ▶ For each SD of  $x$  that we are away from the  $x$  mean, we expect to be  $r$  SDs of  $y$  away from the  $y$  mean.
- ▶ Because  $r$  is always between  $-1$  and  $+1$ , each predicted  $y$  is fewer SDs away from its mean than the corresponding  $x$  was, a phenomenon called regression to the mean.
- ▶ The square of the correlation coefficient,  $R^2$ , gives us the fraction of the variation of the response accounted for by the regression model. The remaining  $1 - R^2$  of the variation is left in the residuals.

The residuals also reveal how well the model works:

- ▶ If a plot of residuals against predicted values shows a pattern, we should re-examine the data to see why.
- ▶ The standard deviation of the residuals,  $s_e$ , quantifies the amount of scatter around the line.

Of course, the linear model makes no sense unless the **Linearity Assumption** is satisfied. We check the **Straight Enough Condition** and **Outlier Condition** with a scatterplot, as we did for correlation, and also with a plot of residuals against either the  $x$  or the predicted values. For the standard deviation of the residuals to make sense as a summary, we have to make the **Equal Variance Assumption**. We check it by looking at both the original scatterplot and the residual plot for the **Does the Plot Thicken? Condition**.

## Terms

Model	172. An equation or formula that simplifies and represents reality.
Linear model	172. A linear model is an equation of a line. To interpret a linear model, we need to know the variables (along with their W's) and their units.
Predicted value	172. The value of $\hat{y}$ found for a given $x$ -value in the data. A predicted value is found by substituting the $x$ -value in the regression equation. The predicted values are the values on the fitted line; the points $(x, \hat{y})$ all lie exactly on the fitted line.
Residuals	172. Residuals are the differences between data values and the corresponding values predicted by the regression model—or, more generally, values predicted by any model.
	Residual = observed value – predicted value = $e = y - \hat{y}$
Least squares	172. The least squares criterion specifies the unique line that minimizes the variance of the residuals or, equivalently, the sum of the squared residuals.
Regression to the mean	174. Because the correlation is always less than 1.0 in magnitude, each predicted $\hat{y}$ tends to be fewer standard deviations from its mean than its corresponding $x$ was from its mean. This is called regression to the mean.
Regression line	174. The particular linear equation
Line of best fit	$\hat{y} = b_0 + b_1x$

that satisfies the least squares criterion is called the least squares regression line. Casually, we often just call it the regression line, or the line of best fit.

**Slope** 176. The slope,  $b_1$ , gives a value in “ $y$ -units *per*  $x$ -unit.” Changes of one unit in  $x$  are associated with changes of  $b_1$  units in predicted values of  $y$ . The slope can be found by

$$b_1 = \frac{rs_y}{s_x}.$$

**Intercept** 176. The intercept,  $b_0$ , gives a starting value in  $y$ -units. It’s the  $\hat{y}$ -value when  $x$  is 0. You can find it from  $b_0 = \bar{y} - b_1\bar{x}$ .

$s_e$  181. The standard deviation of the residuals is found by  $s_e = \sqrt{\frac{\sum e^2}{n-2}}$ . When the assumptions and conditions are met, the residuals can be well described by using this standard deviation and the 68–95–99.7 Rule.

$R^2$

- ▶ 182.  $R^2$  is the square of the correlation between  $y$  and  $x$ .
- ▶  $R^2$  gives the fraction of the variability of  $y$  accounted for by the least squares linear regression on  $x$ .
- ▶  $R^2$  is an overall measure of how successful the regression is in linearly relating  $y$  to  $x$ .

## Skills



- ▶ Be able to identify response ( $y$ ) and explanatory ( $x$ ) variables in context.
- ▶ Understand how a linear equation summarizes the relationship between two variables.
- ▶ Recognize when a regression should be used to summarize a linear relationship between two quantitative variables.
- ▶ Be able to judge whether the slope of a regression makes sense.
- ▶ Know how to examine your data for violations of the **Straight Enough Condition** that would make it inappropriate to compute a regression.
- ▶ Understand that the least squares slope is easily affected by extreme values.
- ▶ Know that residuals are the differences between the data values and the corresponding values predicted by the line and that the *least squares criterion* finds the line that minimizes the sum of the squared residuals.
- ▶ Know how to use a plot of residuals against predicted values to check the **Straight Enough Condition**, the **Does the Plot Thicken? Condition**, and the **Outlier Condition**.
- ▶ Understand that the standard deviation of the residuals,  $s_e$ , measures variability around the line. A large  $s_e$  means the points are widely scattered; a small  $s_e$  means they lie close to the line.



- ▶ Know how to find a regression equation from the summary statistics for each variable and the correlation between the variables.
- ▶ Know how to find a regression equation using your statistics software and how to find the slope and intercept values in the regression output table.
- ▶ Know how to use regression to predict a value of  $y$  for a given  $x$ .
- ▶ Know how to compute the residual for each data value and how to display the residuals.



- ▶ Be able to write a sentence explaining what a linear equation says about the relationship between  $y$  and  $x$ , basing it on the fact that the slope is given in  $y$ -units *per*  $x$ -unit.
- ▶ Understand how the correlation coefficient and the regression slope are related. Know how  $R^2$  describes how much of the variation in  $y$  is accounted for by its linear relationship with  $x$ .
- ▶ Be able to describe a prediction made from a regression equation, relating the predicted value to the specified  $x$ -value.
- ▶ Be able to write a sentence interpreting  $s_e$  as representing typical errors in predictions—the amounts by which actual  $y$ -values differ from the  $\hat{y}$ 's estimated by the model.

## REGRESSION ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:

**AS** Finding Least Squares

**Lines.** We almost always use technology to find regressions. Practice now—just in time for the exercises.

*R squared*

*Standard dev of residuals ( $s_e$ )*

*The “dependent,” response, or y-variable*

Dependent variable is: Total Fat				
R squared = 69.0%				
s = 9.277				
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	6.83077	2.664	2.56	0.0158
Protein	0.971381	0.1209	8.04	≤0.0001

*The “independent,” predictor, or x-variable*

*The slope*

*The intercept*

*We'll deal with all of these later in the book. You may ignore them for now.*

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labeled with the name of the x-variable, and the intercept is labeled “Intercept” or “Constant.” So the regression equation shown here is

$$\widehat{Fat} = 6.83077 + 0.971381 Protein.$$

It is not unusual for statistics packages to give many more digits of the estimated slope and intercept than could possibly be estimated from the data. (The original data were reported to the nearest gram.) Ordinarily, you should round most of the reported numbers to one digit more than the precision of the data, and the slope to two. We will learn about the other numbers in the regression table later in the book. For now, all you need to be able to do is find the coefficients, the  $s_e$ , and the  $R^2$  value.

## EXERCISES

- Cereals.** For many people, breakfast cereal is an important source of fiber in their diets. Cereals also contain potassium, a mineral shown to be associated with maintaining a healthy blood pressure. An analysis of the amount of fiber (in grams) and the potassium content (in milligrams) in servings of 77 breakfast cereals produced the regression model  $\widehat{Potassium} = 38 + 27Fiber$ . If your cereal provides 9 grams of fiber per serving, how much potassium does the model estimate you will get?
- Horsepower.** In Chapter 7's Exercise 33 we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model  $\widehat{mpg} = 46.87 - 0.084HP$ . If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?
- More cereal.** Exercise 1 describes a regression model that estimates a cereal's potassium content from the amount of fiber it contains. In this context, what does it mean to say that a cereal has a negative residual?
- Horsepower, again.** Exercise 2 describes a regression model that uses a car's horsepower to estimate its fuel economy. In this context, what does it mean to say that a certain car has a positive residual?
- Another bowl.** In Exercise 1, the regression model  $\widehat{Potassium} = 38 + 27Fiber$  relates fiber (in grams) and potassium content (in milligrams) in servings of breakfast cereals. Explain what the slope means.
- More horsepower.** In Exercise 2, the regression model  $\widehat{mpg} = 46.87 - 0.084HP$  relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

## REGRESSION ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables may differ slightly from one package to another, but all are essentially the same—and all include much more than we need to know for now. Every computer regression table includes a section that looks something like this:

**A S** Finding Least Squares

**Lines.** We almost always use technology to find regressions. Practice now—just in time for the exercises.

*R squared*

*Standard dev of residuals ( $s_e$ )*

*The “dependent,” response, or y-variable*

Dependent variable is: Total Fat				
R squared = 69.0%				
s = 9.277				
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	6.83077	2.664	2.56	0.0158
Protein	0.971381	0.1209	8.04	≤0.0001

*The “independent,” predictor, or x-variable*

*The slope*

*The intercept*

*We'll deal with all of these later in the book. You may ignore them for now.*

The slope and intercept coefficient are given in a table such as this one. Usually the slope is labeled with the name of the x-variable, and the intercept is labeled “Intercept” or “Constant.” So the regression equation shown here is

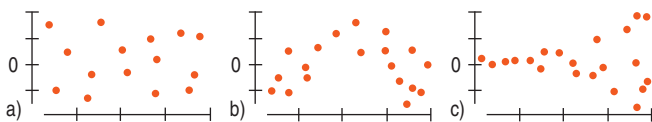
$$\widehat{Fat} = 6.83077 + 0.971381 Protein.$$

It is not unusual for statistics packages to give many more digits of the estimated slope and intercept than could possibly be estimated from the data. (The original data were reported to the nearest gram.) Ordinarily, you should round most of the reported numbers to one digit more than the precision of the data, and the slope to two. We will learn about the other numbers in the regression table later in the book. For now, all you need to be able to do is find the coefficients, the  $s_e$ , and the  $R^2$  value.

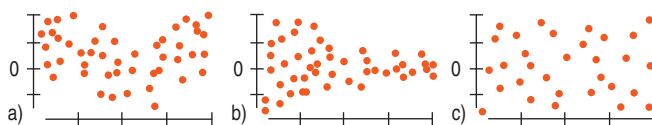
## EXERCISES

- Cereals.** For many people, breakfast cereal is an important source of fiber in their diets. Cereals also contain potassium, a mineral shown to be associated with maintaining a healthy blood pressure. An analysis of the amount of fiber (in grams) and the potassium content (in milligrams) in servings of 77 breakfast cereals produced the regression model  $\widehat{Potassium} = 38 + 27Fiber$ . If your cereal provides 9 grams of fiber per serving, how much potassium does the model estimate you will get?
- Horsepower.** In Chapter 7's Exercise 33 we examined the relationship between the fuel economy (mpg) and horsepower for 15 models of cars. Further analysis produces the regression model  $\widehat{mpg} = 46.87 - 0.084HP$ . If the car you are thinking of buying has a 200-horsepower engine, what does this model suggest your gas mileage would be?
- More cereal.** Exercise 1 describes a regression model that estimates a cereal's potassium content from the amount of fiber it contains. In this context, what does it mean to say that a cereal has a negative residual?
- Horsepower, again.** Exercise 2 describes a regression model that uses a car's horsepower to estimate its fuel economy. In this context, what does it mean to say that a certain car has a positive residual?
- Another bowl.** In Exercise 1, the regression model  $\widehat{Potassium} = 38 + 27Fiber$  relates fiber (in grams) and potassium content (in milligrams) in servings of breakfast cereals. Explain what the slope means.
- More horsepower.** In Exercise 2, the regression model  $\widehat{mpg} = 46.87 - 0.084HP$  relates cars' horsepower to their fuel economy (in mpg). Explain what the slope means.

7. **Cereal again.** The correlation between a cereal's fiber and potassium contents is  $r = 0.903$ . What fraction of the variability in potassium is accounted for by the amount of fiber that servings contain?
8. **Another car.** The correlation between a car's horsepower and its fuel economy (in mpg) is  $r = -0.869$ . What fraction of the variability in fuel economy is accounted for by the horsepower?
9. **Last bowl!** For Exercise 1's regression model predicting potassium content (in milligrams) from the amount of fiber (in grams) in breakfast cereals,  $s_e = 30.77$ . Explain in this context what that means.
10. **Last tank!** For Exercise 2's regression model predicting fuel economy (in mpg) from the car's horsepower,  $s_e = 3.287$ . Explain in this context what that means.
11. **Residuals.** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



12. **Residuals.** Tell what each of the residual plots below indicates about the appropriateness of the linear model that was fit to the data.



13. **What slope?** If you create a regression model for predicting the *Weight* of a car (in pounds) from its *Length* (in feet), is the slope most likely to be 3, 30, 300, or 3000? Explain.
14. **What slope?** If you create a regression model for estimating the *Height* of a pine tree (in feet) based on the *Circumference* of its trunk (in inches), is the slope most likely to be 0.1, 1, 10, or 100? Explain.
15. **Real estate.** A random sample of records of sales of homes from Feb. 15 to Apr. 30, 1993, from the files maintained by the Albuquerque Board of Realtors gives the *Price* and *Size* (in square feet) of 117 homes. A regression to predict *Price* (in thousands of dollars) from *Size* has an  $R$ -squared of 71.4%. The residuals plot indicated that a linear model is appropriate.

- What are the variables and units in this regression?
- What units does the slope have?
- Do you think the slope is positive or negative? Explain.

- T 16. **Roller coaster.** People who responded to a July 2004 Discovery Channel poll named the 10 best roller coasters in the United States. A table in the last chapter's exercises shows the length of the initial drop (in feet) and the duration of the ride (in seconds). A regression to predict *Duration* from *Drop* has  $R^2 = 12.4\%$ .
- What are the variables and units in this regression?
  - What units does the slope have?
  - Do you think the slope is positive or negative? Explain.

17. **Real estate again.** The regression of *Price* on *Size* of homes in Albuquerque had  $R^2 = 71.4\%$ , as described in Exercise 15. Write a sentence (in context, of course) summarizing what the  $R^2$  says about this regression.

- T 18. **Coasters again.** Exercise 16 examined the association between the *Duration* of a roller coaster ride and the height of its initial *Drop*, reporting that  $R^2 = 12.4\%$ . Write a sentence (in context, of course) summarizing what the  $R^2$  says about this regression.

19. **Real estate redux.** The regression of *Price* on *Size* of homes in Albuquerque had  $R^2 = 71.4\%$ , as described in Exercise 15.

- What is the correlation between *Size* and *Price*? Explain why you chose the sign (+ or -) you did.
- What would you predict about the *Price* of a home 1 standard deviation above average in *Size*?
- What would you predict about the *Price* of a home 2 standard deviations below average in *Size*?

- T 20. **Another ride.** The regression of *Duration* of a roller coaster ride on the height of its initial *Drop*, described in Exercise 16, had  $R^2 = 12.4\%$ .

- What is the correlation between *Drop* and *Duration*? Explain why you chose the sign (+ or -) you did.
- What would you predict about the *Duration* of the ride on a coaster whose initial *Drop* was 1 standard deviation below the mean *Drop*?
- What would you predict about the *Duration* of the ride on a coaster whose initial *Drop* was 3 standard deviations above the mean *Drop*?

21. **More real estate.** Consider the Albuquerque home sales from Exercise 15 again. The regression analysis gives the model  $\widehat{Price} = 47.82 + 0.061 \text{ Size}$ .
- Explain what the slope of the line says about housing prices and house size.
  - What price would you predict for a 3000-square-foot house in this market?
  - A real estate agent shows a potential buyer a 1200-square-foot home, saying that the asking price is \$6000 less than what one would expect to pay for a house of this size. What is the asking price, and what is the \$6000 called?

- T 22. **Last ride.** Consider the roller coasters described in Exercise 16 again. The regression analysis gives the model  $\widehat{Duration} = 91.033 + 0.242 \text{ Drop}$ .

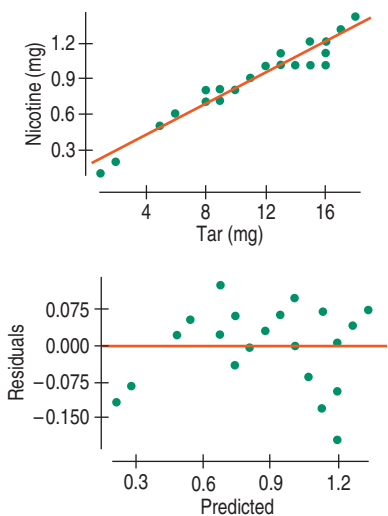
- Explain what the slope of the line says about how long a roller coaster ride may last and the height of the coaster.
- A new roller coaster advertises an initial drop of 200 feet. How long would you predict the rides last?
- Another coaster with a 150-foot initial drop advertises a 2-minute ride. Is this longer or shorter than you'd expect? By how much? What's that called?

23. **Misinterpretations.** A Biology student who created a regression model to use a bird's *Height* when perched for predicting its *Wingspan* made these two statements. Assuming the calculations were done correctly, explain what is wrong with each interpretation.
- My  $R^2$  of 93% shows that this linear model is appropriate.
  - A bird 10 inches tall will have a wingspan of 17 inches.

24. **More misinterpretations.** A Sociology student investigated the association between a country's *Literacy Rate* and *Life Expectancy*, then drew the conclusions listed below. Explain why each statement is incorrect. (Assume that all the calculations were done properly.)
- The *Literacy Rate* determines 64% of the *Life Expectancy* for a country.
  - The slope of the line shows that an increase of 5% in *Literacy Rate* will produce a 2-year improvement in *Life Expectancy*.
25. **ESP.** People who claim to "have ESP" participate in a screening test in which they have to guess which of several images someone is thinking of. You and a friend both took the test. You scored 2 standard deviations above the mean, and your friend scored 1 standard deviation below the mean. The researchers offer everyone the opportunity to take a retest.
- Should you choose to take this retest? Explain.
  - Now explain to your friend what his decision should be and why.

26. **SI jinx.** Players in any sport who are having great seasons, turning in performances that are much better than anyone might have anticipated, often are pictured on the cover of *Sports Illustrated*. Frequently, their performances then falter somewhat, leading some athletes to believe in a "Sports Illustrated jinx." Similarly, it is common for phenomenal rookies to have less stellar second seasons—the so-called "sophomore slump." While fans, athletes, and analysts have proposed many theories about what leads to such declines, a statistician might offer a simpler (statistical) explanation. Explain.

- T 27. **Cigarettes.** Is the nicotine content of a cigarette related to the "tars"? A collection of data (in milligrams) on 29 cigarettes produced the scatterplot, residuals plot, and regression analysis shown:

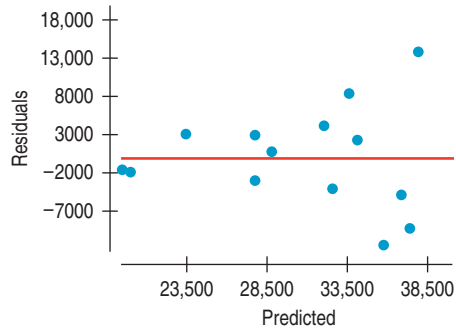
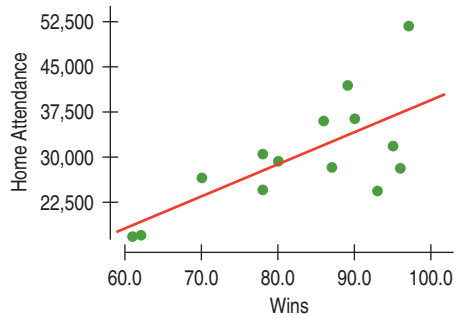


Dependent variable is: nicotine  
 R squared = 92.4%

Variable	Coefficient
Constant	0.154030
Tar	0.065052

- Do you think a linear model is appropriate here? Explain.
- Explain the meaning of  $R^2$  in this context.

- T 28. **Attendance 2006.** In the previous chapter you looked at the relationship between the number of wins by American League baseball teams and the average attendance at their home games for the 2006 season. Here are the scatterplot, the residuals plot, and part of the regression analysis:



Dependent variable is: Home Attendance  
 R squared = 48.5%

Variable	Coefficient
Constant	-14364.5
Wins	538.915

- Do you think a linear model is appropriate here? Explain.
- Interpret the meaning of  $R^2$  in this context.
- Do the residuals show any pattern worth remarking on?
- The point in the upper right of the plots is the New York Yankees. What can you say about the residual for the Yankees?

- T 29. **Another cigarette.** Consider again the regression of *Nicotine* content on *Tar* (both in milligrams) for the cigarettes examined in Exercise 27.
- What is the correlation between *Tar* and *Nicotine*?
  - What would you predict about the average *Nicotine* content of cigarettes that are 2 standard deviations below average in *Tar* content?
  - If a cigarette is 1 standard deviation above average in *Nicotine* content, what do you suspect is true about its *Tar* content?

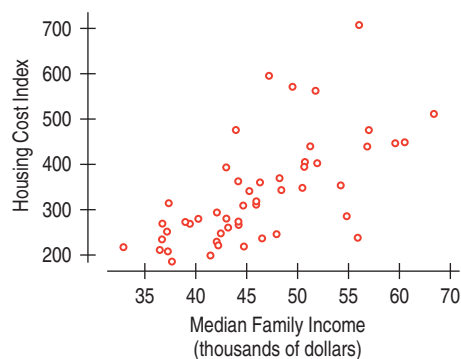
- T 30. **Second inning 2006.** Consider again the regression of *Average Attendance* on *Wins* for the baseball teams examined in Exercise 28.

- What is the correlation between *Wins* and *Average Attendance*?
- What would you predict about the *Average Attendance* for a team that is 2 standard deviations above average in *Wins*?
- If a team is 1 standard deviation below average in attendance, what would you predict about the number of games the team has won?

- T 31. Last cigarette.** Take another look at the regression analysis of tar and nicotine content of the cigarettes in Exercise 27.
- Write the equation of the regression line.
  - Estimate the *Nicotine* content of cigarettes with 4 milligrams of *Tar*.
  - Interpret the meaning of the slope of the regression line in this context.
  - What does the  $y$ -intercept mean?
  - If a new brand of cigarette contains 7 milligrams of tar and a nicotine level whose residual is  $-0.5$  mg, what is the nicotine content?

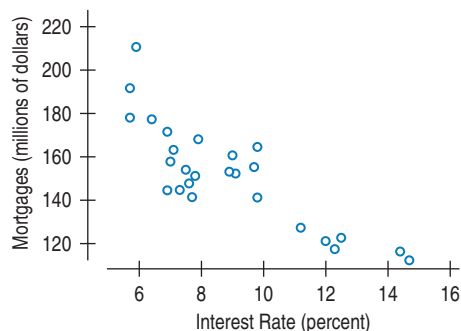
- T 32. Last inning 2006.** Refer again to the regression analysis for average attendance and games won by American League baseball teams, seen in Exercise 28.
- Write the equation of the regression line.
  - Estimate the *Average Attendance* for a team with 50 *Wins*.
  - Interpret the meaning of the slope of the regression line in this context.
  - In general, what would a negative residual mean in this context?
  - The St. Louis Cardinals, the 2006 World Champions, are not included in these data because they are a National League team. During the 2006 regular season, the Cardinals won 83 games and averaged 42,588 fans at their home games. Calculate the residual for this team, and explain what it means.

- T 33. Income and housing revisited.** In Chapter 7, Exercise 31, we learned that the Office of Federal Housing Enterprise Oversight (OFHEO) collects data on various aspects of housing costs around the United States. Here's a scatterplot (by state) of the *Housing Cost Index* (HCI) versus the *Median Family Income* (MFI) for the 50 states. The correlation is  $r = 0.65$ . The mean HCI is 338.2, with a standard deviation of 116.55. The mean MFI is \$46,234, with a standard deviation of \$7072.47.

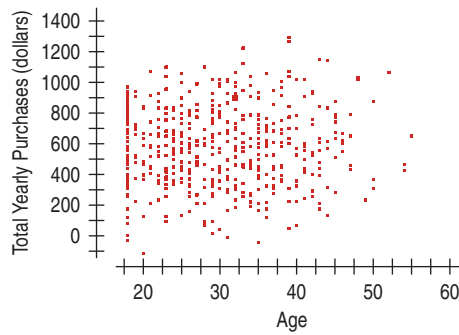


- Is a regression analysis appropriate? Explain.
- What is the equation that predicts Housing Cost Index from median family income?
- For a state with MFI = \$44,993, what would be the predicted HCI?
- Washington, DC, has an MFI of \$44,993 and an HCI of 548.02. How far off is the prediction in b) from the actual HCI?
- If we standardized both variables, what would be the regression equation that predicts standardized HCI from standardized MFI?
- If we standardized both variables, what would be the regression equation that predicts standardized MFI from standardized HCI?

- 34. Interest rates and mortgages again.** In Chapter 7, Exercise 32, we saw a plot of total mortgages in the United States (in millions of 2005 dollars) versus the interest rate at various times over the past 26 years. The correlation is  $r = -0.84$ . The mean mortgage amount is \$151.9 million and the mean interest rate is 8.88%. The standard deviations are \$23.86 million for mortgage amounts and 2.58% for the interest rates.



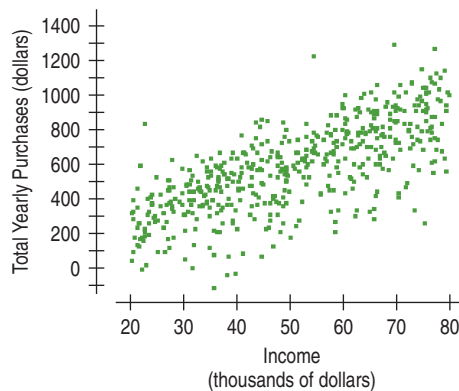
- Is a regression model appropriate for predicting mortgage amount from interest rates? Explain.
  - What is the equation that predicts mortgage amount from interest rates?
  - What would you predict the mortgage amount would be if the interest rates climbed to 20%?
  - Do you have any reservations about your prediction in part c)?
  - If we standardized both variables, what would be the regression equation that predicts standardized mortgage amount from standardized interest rates?
  - If we standardized both variables, what would be the regression equation that predicts standardized interest rates from standardized mortgage amount?
- 35. Online clothes.** An online clothing retailer keeps track of its customers' purchases. For those customers who signed up for the company's credit card, the company also has information on the customer's *Age* and *Income*. A random sample of 500 of these customers shows the following scatterplot of *Total Yearly Purchases* by *Age*:



The correlation between *Total Yearly Purchases* and *Age* is  $r = 0.037$ . Summary statistics for the two variables are:

	Mean	SD
Age	29.67 yrs	8.51 yrs
Total Yearly Purchase	\$572.52	\$253.62

- What is the linear regression equation for predicting *Total Yearly Purchase* from *Age*?
  - Do the assumptions and conditions for regression appear to be met?
  - What is the predicted average *Total Yearly Purchase* for an 18-year-old? For a 50-year-old?
  - What percent of the variability in *Total Yearly Purchases* is accounted for by this model?
  - Do you think the regression might be a useful one for the company? Explain.
36. **Online clothes II.** For the online clothing retailer discussed in the previous problem, the scatterplot of *Total Yearly Purchases* by *Income* shows



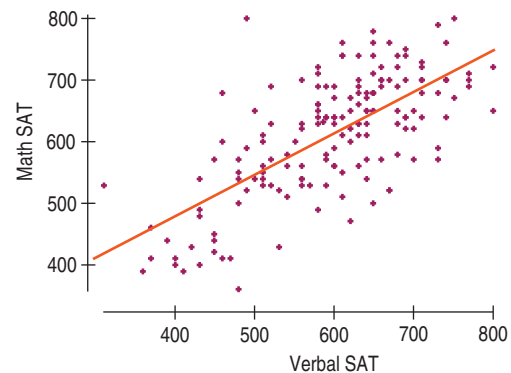
The correlation between *Total Yearly Purchases* and *Income* is 0.722. Summary statistics for the two variables are:

	Mean	SD
Income	\$50,343.40	\$16,952.50
Total Yearly Purchase	\$572.52	\$253.62

- What is the linear regression equation for predicting *Total Yearly Purchase* from *Income*?
- Do the assumptions and conditions for regression appear to be met?

- What is the predicted average *Total Yearly Purchase* for someone with a yearly *Income* of \$20,000? For someone with an annual *Income* of \$80,000?
- What percent of the variability in *Total Yearly Purchases* is accounted for by this model?
- Do you think the regression might be a useful one for the company? Comment.

- T 37. SAT scores.** The SAT is a test often used as part of an application to college. SAT scores are between 200 and 800, but have no units. Tests are given in both Math and Verbal areas. Doing the SAT-Math problems also involves the ability to read and understand the questions, but can a person's verbal score be used to predict the math score? Verbal and math SAT scores of a high school graduating class are displayed in the scatterplot, with the regression line added.



- Describe the relationship.
  - Are there any students whose scores do not seem to fit the overall pattern?
  - For these data,  $r = 0.685$ . Interpret this statistic.
  - These verbal scores averaged 596.3, with a standard deviation of 99.5, and the math scores averaged 612.2, with a standard deviation of 96.1. Write the equation of the regression line.
  - Interpret the slope of this line.
  - Predict the math score of a student with a verbal score of 500.
  - Every year some student scores a perfect 1600. Based on this model, what would be that student's Math score residual?
38. **Success in college.** Colleges use SAT scores in the admissions process because they believe these scores provide some insight into how a high school student will perform at the college level. Suppose the entering freshmen at a certain college have mean combined *SAT Scores* of 1833, with a standard deviation of 123. In the first semester these students attained a mean *GPA* of 2.66, with a standard deviation of 0.56. A scatterplot showed the association to be reasonably linear, and the correlation between *SAT score* and *GPA* was 0.47.
- Write the equation of the regression line.
  - Explain what the  $y$ -intercept of the regression line indicates.
  - Interpret the slope of the regression line.
  - Predict the *GPA* of a freshman who scored a combined 2100.



- e) Based upon these statistics, how effective do you think SAT scores would be in predicting academic success during the first semester of the freshman year at this college? Explain.
- f) As a student, would you rather have a positive or a negative residual in this context? Explain.

39. **SAT, take 2.** Suppose we wanted to use SAT math scores to estimate verbal scores based on the information in Exercise 37.

- What is the correlation?
- Write the equation of the line of regression predicting verbal scores from math scores.
- In general, what would a positive residual mean in this context?
- A person tells you her math score was 500. Predict her verbal score.
- Using that predicted verbal score and the equation you created in Exercise 37, predict her math score.
- Why doesn't the result in part e) come out to 500?

40. **Success, part 2.** Based on the statistics for college freshmen given in Exercise 38, what SAT score might be expected among freshmen who attained a first-semester GPA of 3.0?

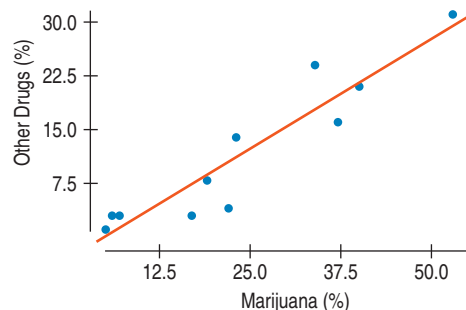
**T** 41. **Used cars 2007.** Classified ads in the *Ithaca Journal* offered several used Toyota Corollas for sale. Listed below are the ages of the cars and the advertised prices.

Age (yr)	Price Advertised (\$)
1	13,990
1	13,495
3	12,999
4	9500
4	10,495
5	8995
5	9495
6	6999
7	6950
7	7850
8	6999
8	5995
10	4950
10	4495
13	2850

- Make a scatterplot for these data.
- Describe the association between *Age* and *Price* of a used Corolla.
- Do you think a linear model is appropriate?
- Computer software says that  $R^2 = 94.4\%$ . What is the correlation between *Age* and *Price*?
- Explain the meaning of  $R^2$  in this context.
- Why doesn't this model explain 100% of the variability in the price of a used Corolla?

**T** 42. **Drug abuse.** In the exercises of the last chapter you examined results of a survey conducted in the United States and 10 countries of Western Europe to determine the

percentage of teenagers who had used marijuana and other drugs. Below is the scatterplot. Summary statistics showed that the mean percent that had used marijuana was 23.9%, with a standard deviation of 15.6%. An average of 11.6% of teens had used other drugs, with a standard deviation of 10.2%.



- Do you think a linear model is appropriate? Explain.
- For this regression,  $R^2$  is 87.3%. Interpret this statistic in this context.
- Write the equation you would use to estimate the percentage of teens who use other drugs from the percentage who have used marijuana.
- Explain in context what the slope of this line means.
- Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs?

**T** 43. **More used cars 2007.** Use the advertised prices for Toyota Corollas given in Exercise 41 to create a linear model for the relationship between a car's *Age* and its *Price*.

- Find the equation of the regression line.
- Explain the meaning of the slope of the line.
- Explain the meaning of the  $y$ -intercept of the line.
- If you want to sell a 7-year-old Corolla, what price seems appropriate?
- You have a chance to buy one of two cars. They are about the same age and appear to be in equally good condition. Would you rather buy the one with a positive residual or the one with a negative residual? Explain.
- You see a "For Sale" sign on a 10-year-old Corolla stating the asking price as \$3500. What is the residual?
- Would this regression model be useful in establishing a fair price for a 20-year-old car? Explain.

**T** 44. **Birthrates 2005.** The table shows the number of live births per 1000 women aged 15–44 years in the United States, starting in 1965. (National Center for Health Statistics, [www.cdc.gov/nchs/](http://www.cdc.gov/nchs/))

Year	1965	1970	1975	1980	1985	1990	1995	2000	2005
Rate	19.4	18.4	14.8	15.9	15.6	16.4	14.8	14.4	14.0

- Make a scatterplot and describe the general trend in *Birthrates*. (Enter *Year* as years since 1900: 65, 70, 75, etc.)
- Find the equation of the regression line.
- Check to see if the line is an appropriate model. Explain.

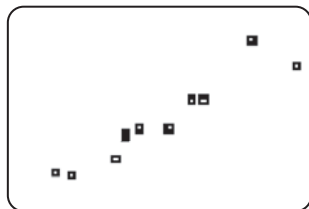
- d) Interpret the slope of the line.
- e) The table gives rates only at 5-year intervals. Estimate what the rate was in 1978.
- f) In 1978 the birthrate was actually 15.0. How close did your model come?
- g) Predict what the *Birthrate* will be in 2010. Comment on your faith in this prediction.
- h) Predict the *Birthrate* for 2025. Comment on your faith in this prediction.

45. **Burgers.** In the last chapter, you examined the association between the amounts of *Fat* and *Calories* in fast-food hamburgers. Here are the data:

Fat (g)	19	31	34	35	39	39	43
Calories	410	580	590	570	640	680	660

- a) Create a scatterplot of *Calories* vs. *Fat*.
  - b) Interpret the value of  $R^2$  in this context.
  - c) Write the equation of the line of regression.
  - d) Use the residuals plot to explain whether your linear model is appropriate.
  - e) Explain the meaning of the  $y$ -intercept of the line.
  - f) Explain the meaning of the slope of the line.
  - g) A new burger containing 28 grams of fat is introduced. According to this model, its residual for calories is +33. How many calories does the burger have?
46. **Chicken.** Chicken sandwiches are often advertised as a healthier alternative to beef because many are lower in fat. Tests on 11 brands of fast-food chicken sandwiches produced the following summary statistics and scatterplot from a graphing calculator:

	<b>Fat (g)</b>	<b>Calories</b>
<b>Mean</b>	20.6	472.7
<b>St. Dev.</b>	9.8	144.2
<b>Correlation</b>	0.947	



- a) Do you think a linear model is appropriate in this situation?
- b) Describe the strength of this association.
- c) Write the equation of the regression line to estimate calories from the fat content.
- d) Explain the meaning of the slope.
- e) Explain the meaning of the  $y$ -intercept.
- f) What does it mean if a certain sandwich has a negative residual?

47. **A second helping of burgers.** In Exercise 45 you created a model that can estimate the number of *Calories* in a burger when the *Fat* content is known.

- a) Explain why you cannot use that model to estimate the fat content of a burger with 600 calories.
- b) Using an appropriate model, estimate the fat content of a burger with 600 calories.

48. **A second helping of chicken.** In Exercise 46 you created a model to estimate the number of *Calories* in a chicken sandwich when you know the *Fat*.

- a) Explain why you cannot use that model to estimate the fat content of a 400-calorie sandwich.
- b) Make that estimate using an appropriate model.

T 49. **Body fat.** It is difficult to determine a person's body fat percentage accurately without immersing him or her in water. Researchers hoping to find ways to make a good estimate immersed 20 male subjects, then measured their waists and recorded their weights.

Waist (in.)	Weight (lb)	Body Fat (%)	Waist (in.)	Weight (lb)	Body Fat (%)
32	175	6	33	188	10
36	181	21	40	240	20
38	200	15	36	175	22
33	159	6	32	168	9
39	196	22	44	246	38
40	192	31	33	160	10
41	205	32	41	215	27
35	173	21	34	159	12
38	187	25	34	146	10
38	188	30	44	219	28

- a) Create a model to predict %*Body Fat* from *Weight*.
- b) Do you think a linear model is appropriate? Explain.
- c) Interpret the slope of your model.
- d) Is your model likely to make reliable estimates? Explain.
- e) What is the residual for a person who weighs 190 pounds and has 21% body fat?

T 50. **Body fat again.** Would a model that uses the person's *Waist* size be able to predict the %*Body Fat* more accurately than one that uses *Weight*? Using the data in Exercise 49, create and analyze that model.

T 51. **Heptathlon 2004.** We discussed the women's 2004 Olympic heptathlon in Chapter 6. The table on the next page shows the results from the high jump, 800-meter run, and long jump for the 26 women who successfully completed all three events in the 2004 Olympics.

Name	Country	High Jump (m)	800-m (sec)	Long Jump (m)
Carolina Klüft	SWE	1.91	134.15	6.51
Austra Skujytė	LIT	1.76	135.92	6.30
Kelly Sotherton	GBR	1.85	132.27	6.51
Shelia Burrell	USA	1.70	135.32	6.25
Yelena Prokhorova	RUS	1.79	131.31	6.21
Sonja Kesselschlaeger	GER	1.76	135.21	6.42
Marie Collonville	FRA	1.85	133.62	6.19
Natalya Dobrynska	UKR	1.82	137.01	6.23
Margaret Simpson	GHA	1.79	137.72	6.02
Svetlana Sokolova	RUS	1.70	133.23	5.84
J. J. Shobha	IND	1.67	137.28	6.36
Claudia Tonn	GER	1.82	130.77	6.35
Naide Gomes	POR	1.85	140.05	6.10
Michelle Perry	USA	1.70	133.69	6.02
Aryiro Strataki	GRE	1.79	137.90	5.97
Karin Ruckstuhl	NED	1.85	133.95	5.90
Karin Ertl	GER	1.73	138.68	6.03
Kylie Wheeler	AUS	1.79	137.65	6.36
Janice Josephs	RSA	1.70	138.47	6.21
Tiffany Lott Hogan	USA	1.67	145.10	6.15
Magdalena Szczepanska	POL	1.76	133.08	5.98
Irina Naumenko	KAZ	1.79	134.57	6.16
Yuliya Akulenko	UKR	1.73	142.58	6.02
Soma Biswas	IND	1.70	132.27	5.92
Marsha Mark-Baird	TRI	1.70	141.21	6.22
Michaela Hejnova	CZE	1.70	145.68	5.70

Let's examine the association among these events. Perform a regression to predict high-jump performance from the 800-meter results.

- What is the regression equation? What does the slope mean?
- What percent of the variability in high jumps can be accounted for by differences in 800-m times?
- Do good high jumpers tend to be fast runners? (Be careful—low times are good for running events and high distances are good for jumps.)
- What does the residuals plot reveal about the model?
- Do you think this is a useful model? Would you use it to predict high-jump performance? (Compare the residual standard deviation to the standard deviation of the high jumps.)

**T 52. Heptathlon 2004 again.** We saw the data for the women's 2004 Olympic heptathlon in Exercise 51. Are the two jumping events associated? Perform a regression of the long-jump results on the high-jump results.

- What is the regression equation? What does the slope mean?
- What percentage of the variability in long jumps can be accounted for by high-jump performances?
- Do good high jumpers tend to be good long jumpers?
- What does the residuals plot reveal about the model?
- Do you think this is a useful model? Would you use it to predict long-jump performance? (Compare the residual standard deviation to the standard deviation of the long jumps.)

**53. Least squares.** Consider the four points (10,10), (20,50), (40,20), and (50,80). The least squares line is  $\hat{y} = 7.0 + 1.1x$ . Explain what "least squares" means, using these data as a specific example.

**54. Least squares.** Consider the four points (200,1950), (400,1650), (600,1800), and (800,1600). The least squares line is  $\hat{y} = 1975 - 0.45x$ . Explain what "least squares" means, using these data as a specific example.



## JUST CHECKING

### Answers

1. You should expect the price to be 0.77 standard deviations above the mean.
2. You should expect the size to be  $2(0.77) = 1.54$  standard deviations below the mean.
3. The home is 1.5 standard deviations above the mean in size no matter how size is measured.
4. An increase in home size of 1000 square feet is associated with an increase in price of \$94,454, on average.
5. Units are thousands of dollars per thousand square feet.
6. About \$188,908, on average
7. No. Even if it were positive, no one wants a house with 0 square feet!
8. Negative; that indicates it's priced lower than a typical home of its size.
9. \$280,245
10. \$19,755 (positive!)
11. Differences in the size of houses account for about 59.5% of the variation in the house prices.
12. It's positive. The correlation and the slope have the same sign.
13.  $R^2$  would not change, but the slope would. Slope depends on the units used but correlation doesn't.
14. No, the standard deviation of the residuals is 53.79 thousand dollars. We shouldn't be surprised by any residual smaller than 2 standard deviations, and a residual of \$100,000 is less than  $2(53,790)$ .

# Regression Wisdom



AS

**Activity: Construct a Plot with a Given Slope.** How's your feel for regression lines? Can you make a scatterplot that has a specified slope?

Regression may be the most widely used Statistics method. It is used every day throughout the world to predict customer loyalty, numbers of admissions at hospitals, sales of automobiles, and many other things. Because regression is so widely used, it's also widely abused and misinterpreted. This chapter presents examples of regressions in which things are not quite as simple as they may have seemed at first, and shows how you can still use regression to discover what the data have to say.

## Getting the “Bends”: When the Residuals Aren't Straight

No regression analysis is complete without a display of the residuals to check that the linear model is reasonable. Because the residuals are what is “left over” after the model describes the relationship, they often reveal subtleties that were not clear from a plot of the original data. Sometimes these are additional details that help confirm or refine our understanding. Sometimes they reveal violations of the regression conditions that require our attention.

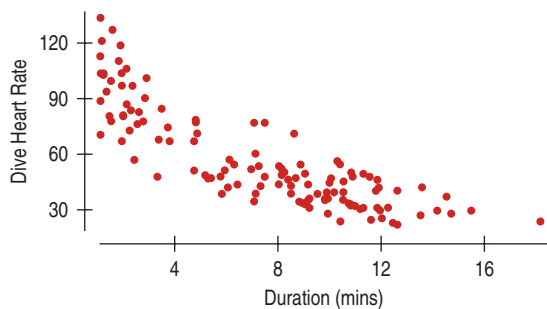
The fundamental assumption in working with a linear model is that the relationship you are modeling is, in fact, linear. That sounds obvious, but when you fit a regression, you can't take it for granted. Often it's hard to tell from the scatterplot you looked at before you fit the regression model. Sometimes you can't see a bend in the relationship until you plot the residuals.

Jessica Meir and Paul Ponganis study emperor penguins at the Scripps Institution of Oceanography's Center for Marine Biotechnology and Biomedicine at the University of California at San Diego. Says Jessica:

*Emperor penguins are the most accomplished divers among birds, making routine dives of 5–12 minutes, with the longest recorded dive over 27 minutes. These birds can also dive to depths of over 500 meters! Since air-breathing animals like penguins must hold their breath while submerged, the duration of any given dive depends on how much oxygen is in the bird's body at the beginning of the dive, how quickly that oxygen gets used,*

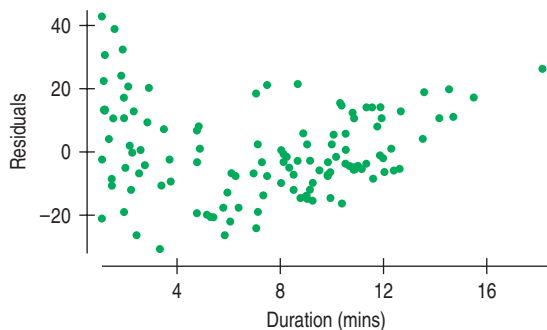
We can't *know* whether the **Linearity Assumption** is true, but we can see if it's *plausible* by checking the **Straight Enough Condition**.

and the lowest level of oxygen the bird can tolerate. The rate of oxygen depletion is primarily determined by the penguin's heart rate. Consequently, studies of heart rates during dives can help us understand how these animals regulate their oxygen consumption in order to make such impressive dives.



**FIGURE 9.1**

The scatterplot of Dive Heart Rate in beats per minute (bpm) vs. Duration (minutes) shows a strong, roughly linear, negative association.



**FIGURE 9.2**

Plotting the residuals against Duration reveals a bend. It was also in the original scatterplot, but here it's easier to see.

The researchers equip emperor penguins with devices that record their heart rates during dives. Here's a scatterplot of the *Dive Heart Rate* (beats per minute) and the *Duration* (minutes) of dives by these high-tech penguins.

The scatterplot looks fairly linear with a moderately strong negative association ( $R^2 = 71.5\%$ ). The linear regression equation

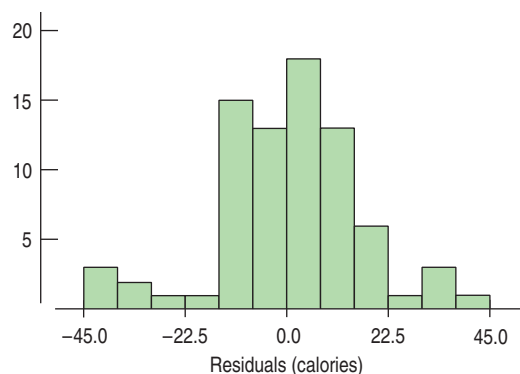
$$\widehat{DiveHeartRate} = 96.9 - 5.47 \text{ Duration}$$

says that for longer dives, the average *Dive Heart Rate* is lower by about 5.47 beats per dive minute, starting from a value of 96.9 beats per minute.

The scatterplot of the residuals against *Duration* holds a surprise. The Linearity Assumption says we should not see a pattern, but instead there's a bend, starting high on the left, dropping down in the middle of the plot, and rising again at the right. Graphs of residuals often reveal patterns such as this that were easy to miss in the original scatterplot.

Now looking back at the original scatterplot, you may see that the scatter of points isn't really straight. There's a slight bend to that plot, but the bend is much easier to see in the residuals. Even though it means rechecking the Straight Enough Condition *after* you find the regression, it's always a good idea to check your scatterplot of the residuals for bends that you might have overlooked in the original scatterplot.

## Sifting Residuals for Groups



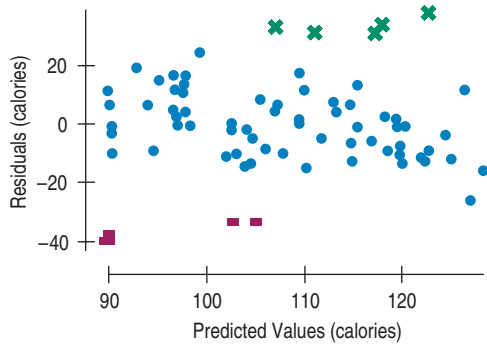
**FIGURE 9.3**

A histogram of the regression residuals shows small modes both above and below the central large mode. These may be worth a second look.

In the Step-By-Step analysis in Chapter 8 to predict *Calories* from *Sugar* content in breakfast cereals, we examined a scatterplot of the residuals. Our first impression was that it had no particular structure—a conclusion that supported using the regression model. But let's look again.

Here's a histogram of the residuals. How would you describe its shape? It looks like there might be small modes on both sides of the central body of the data. One group of cereals seems to stand out as having large negative residuals, with fewer calories than we might have predicted, and another stands out with large positive residuals. The calories in these cereals were underestimated by the model. Whenever we suspect multiple modes, we ask whether they are somehow different.

On the next page is the residual plot, with the points in those modes marked. Now we can see that those two groups stand away from the central pattern in the scatterplot. The high-residual cereals are Just Right Fruit & Nut; Muesli Raisins, Dates & Almonds; Peaches & Pecans; Mueslix Crispy Blend; and Nutri-Grain Almond Raisin. Do these cereals seem to have something in common? They all present themselves as "healthy." This might be surprising, but in fact, "healthy" cereals


**FIGURE 9.4**

A scatterplot of the residuals vs. predicted values for the cereal regression. The green “x” points are cereals whose calorie content is higher than the linear model predicts. The red “-” points show cereals with fewer calories than the model predicts. Is there something special about these cereals?

often contain more fat, and therefore more calories, than we might expect from looking at their sugar content alone.

The low-residual cereals are Puffed Rice, Puffed Wheat, three bran cereals, and Golden Crisps. You might not have grouped these cereals together before. What they have in common is a low calorie count *relative to their sugar content*—even though their sugar contents are quite different.

These observations may not lead us to question the overall linear model, but they do help us understand that other factors may be part of the story. An examination of residuals often leads us to discover groups of observations that are different from the rest.

When we discover that there is more than one group in a regression, we may decide to analyze the groups separately, using a different model for each group. Or we can stick with the original model and simply note that there are groups that are a little different. Either way, the model will be wrong, but useful, so it will improve our understanding of the data.

## Subsets

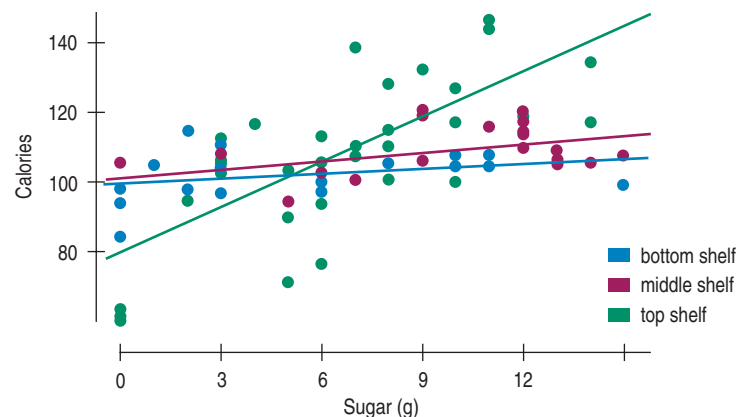
Here’s an important unstated condition for fitting models: **All the data must come from the same population.**

Cereal manufacturers aim cereals at different segments of the market. Supermarkets and cereal manufacturers try to attract different customers by placing different types of cereals on certain shelves. Cereals for kids tend to be on the “kid’s shelf,” at their eye level. Toddlers wouldn’t be likely to grab a box from this shelf and beg, “Mom, can we please get this All-Bran with Extra Fiber?”

Should we take this extra information into account in our analysis? Figure 9.5 shows a scatterplot of *Calories* and *Sugar*, colored according to the shelf on which the cereals were found and with a separate regression line fit for each. The top shelf is clearly different. We might want to report two regressions, one for the top shelf and one for the bottom two shelves.<sup>1</sup>

**FIGURE 9.5**

Calories and Sugar colored according to the shelf on which the cereal was found in a supermarket, with regression lines fit for each shelf individually. Do these data appear homogeneous? That is, do all the cereals seem to be from the same population of cereals? Or are there different kinds of cereals that we might want to consider separately?



## Extrapolation: Reaching Beyond the Data

Linear models give a predicted value for each case in the data. Put a new  $x$ -value into the equation, and it gives a predicted value,  $\hat{y}$ , to go with it. But when the new  $x$ -value lies far from the data we used to build the regression, how trustworthy is the prediction?

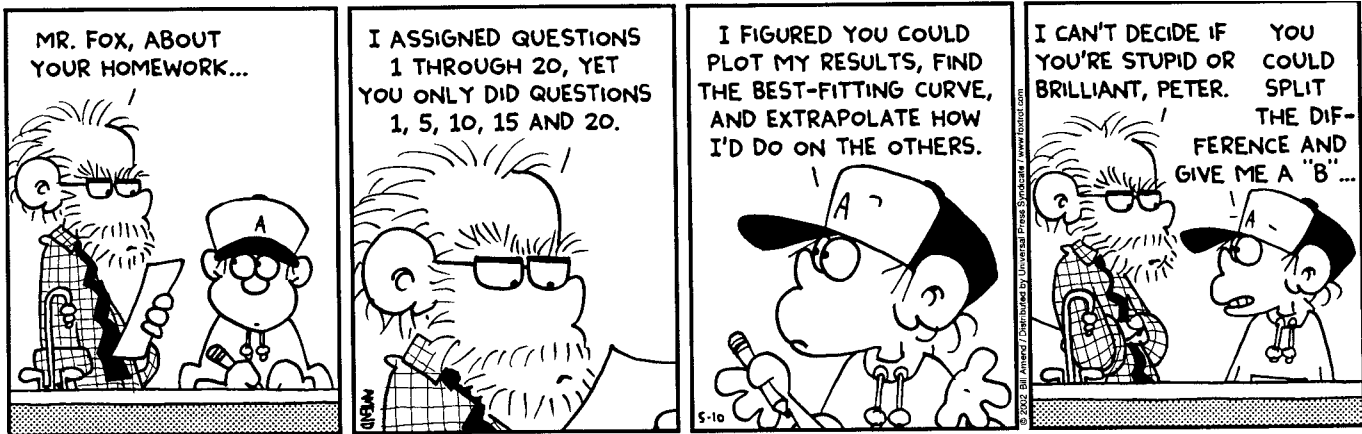
<sup>1</sup> More complex models can take into account both sugar content and shelf information. This kind of *multiple regression* model is a natural extension of the model we’re using here. You can learn about such models in Chapter 29 on the DVD.

**AS** **Case Study: Predicting Manatee Kills.** Can we use regression to predict the number of manatees that will be killed by power boats this year?

*“Prediction is difficult, especially about the future.”*  
 —Niels Bohr, Danish physicist

The simple answer is that the farther the new  $x$ -value is from  $\bar{x}$ , the less trust we should place in the predicted value. Once we venture into new  $x$  territory, such a prediction is called an **extrapolation**. Extrapolations are dubious because they require the very questionable assumption that nothing about the relationship between  $x$  and  $y$  changes even at extreme values of  $x$  and beyond.

Extrapolations can get us into deep trouble. When the  $x$ -variable is *Time*, extrapolation becomes an attempt to peer into the future. People have always wanted to see into the future, and it doesn't take a crystal ball to foresee that they always will. In the past, seers, oracles, and wizards were called on to predict the future. Today mediums, fortune-tellers, and Tarot card readers still find many customers.



FOXTROT © 2002 Bill Amend. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.



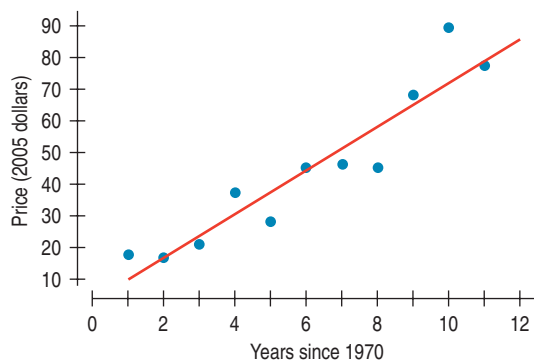
**When the Data Are Years . . .**

... we usually don't enter them as four-digit numbers. Here we used 0 for 1970, 10 for 1980, and so on. Or we may simply enter two digits, using 82 for 1982, for instance. Rescaling years like this often makes calculations easier and equations simpler. We recommend you do it, too. But be careful: If 1982 is 82, then 2004 is 104 (not 4), right?

Those with a more scientific outlook may use a linear model as their digital crystal ball. Linear models are based on the  $x$ -values of the data at hand and cannot be trusted beyond that span. Some physical phenomena do exhibit a kind of “inertia” that allows us to guess that current systematic behavior will continue, but regularity can't be counted on in phenomena such as stock prices, sales figures, hurricane tracks, or public opinion.

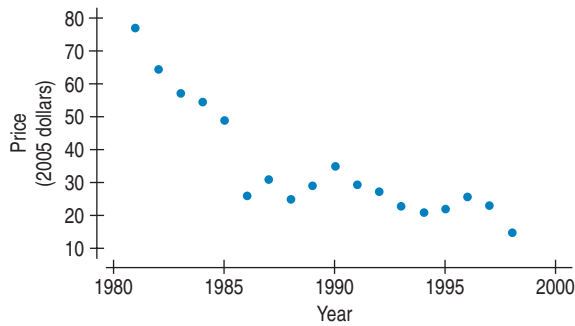
Extrapolating from current trends is so tempting that even professional forecasters make this mistake, and sometimes the errors are striking. In the mid-1970s, oil prices surged and long lines at gas stations were common. In 1970, oil cost about \$17 a barrel (in 2005 dollars)—about what it had cost for 20 years or so. But then, within just a few years, the price surged to over \$40. In 1975, a survey of 15 top econometric forecasting models (built by groups that included Nobel prize-winning economists) found predictions for 1985 oil prices that ranged from \$300 to over \$700 a barrel (in 2005 dollars). How close were these forecasts?

Here's a scatterplot of oil prices from 1972 to 1981 (in 2005 dollars).

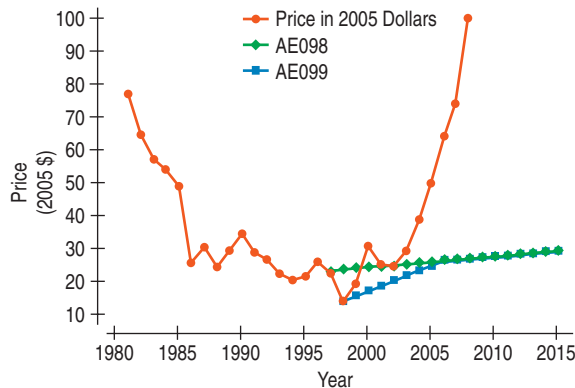


**FIGURE 9.6**  
 The scatterplot shows an average increase in the price of a barrel of oil of over \$7 per year from 1971 to 1982.



**FIGURE 9.7**

This scatterplot of oil prices from 1981 to 1998 shows a fairly constant decrease of about \$3 per barrel per year.

**FIGURE 9.8**

Here are the EIA forecasts with the actual prices from 1981 to 2008. Neither forecast predicted the sharp run-up in the past few years.

The regression model

$$\widehat{\text{Price}} = -0.85 + 7.39 \text{ Years since 1970}$$

says that prices had been going up 7.39 dollars per year, or nearly \$74 in 10 years. If you assume that they would *keep going up*, it's not hard to imagine almost any price you want.

So, how did the forecasters do? Well, in the period from 1982 to 1998 oil prices didn't exactly continue that steady increase. In fact, they went down so much that by 1998, prices (adjusted for inflation) were the lowest they'd been since before World War II.

Not one of the experts' models predicted that.

Of course, these decreases clearly couldn't continue, or oil would be free by now. The Energy Information Administration offered two *different* 20-year forecasts for oil prices after 1998, and both called for relatively modest increases in oil prices. So, how accurate have *these* forecasts been? Here's a timeplot of the EIA's predictions and the actual prices (in 2005 dollars).

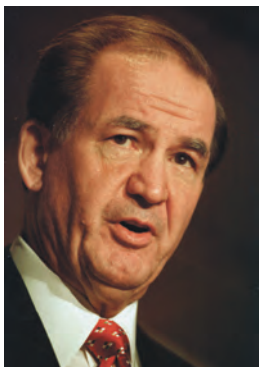
Oops! They seemed to have missed the sharp run-up in oil prices in the past few years.

Where do you think oil prices will go in the next decade? *Your* guess may be as good as anyone's!

Of course, knowing that extrapolation is dangerous doesn't stop people. The temptation to see into the future is hard to resist. So our more realistic advice is this:

*If you must extrapolate into the future, at least don't believe that the prediction will come true.*

## Outliers, Leverage, and Influence



The outcome of the 2000 U.S. presidential election was determined in Florida amid much controversy. The main race was between George W. Bush and Al Gore, but two minor candidates played a significant role. To the political right of the main party candidates was Pat Buchanan, while to the political left was Ralph Nader. Generally, Nader earned more votes than Buchanan throughout the state. We would expect counties with larger vote totals to give more votes to each candidate. Here's a regression relating *Buchanan's* vote totals by county in the state of Florida to *Nader's*:

Dependent variable is: Buchanan

R-squared = 42.8%

Variable	Coefficient
Intercept	50.3
Nader	0.14

The regression model,

$$\widehat{\text{Buchanan}} = 50.3 + 0.14 \text{ Nader},$$

says that, in each county, Buchanan received about 0.14 times (or 14% of) the vote Nader received, starting from a base of 50.3 votes.

This seems like a reasonable regression, with an  $R^2$  of almost 43%. But we've violated all three Rules of Data Analysis by going straight to the regression table without making a picture.

Here's a scatterplot that shows the vote for Buchanan in each county of Florida plotted against the vote for Nader. The striking **outlier** is Palm Beach County.

“Nature is nowhere accustomed more openly to display her secret mysteries than in cases where she shows traces of her workings apart from the beaten path.”

—William Harvey (1657)

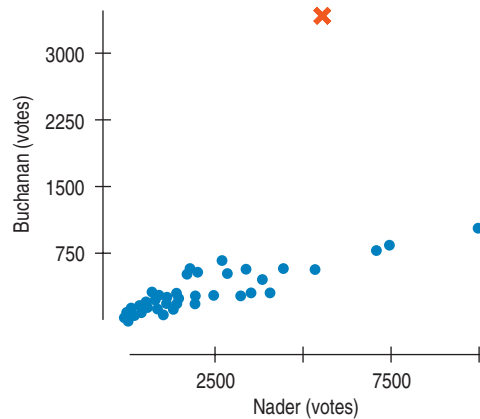


FIGURE 9.9

Votes received by Buchanan against votes for Nader in all Florida counties in the presidential election of 2000. The red “x” point is Palm Beach County, home of the “butterfly ballot.”

The so-called “butterfly ballot,” used only in Palm Beach County, was a source of controversy. It has been claimed that the format of this ballot confused voters so that some who intended to vote for the Democrat, Al Gore, punched the wrong hole next to his name and, as a result, voted for Buchanan.

The scatterplot shows a strong, positive, linear association, and one striking point. With Palm Beach removed from the regression, the  $R^2$  jumps from 42.8% to 82.1% and the slope of the line changes to 0.1, suggesting that Buchanan received only about 10% of the vote that Nader received. With more than 82% of the variability of the Buchanan vote accounted for, the model when Palm Beach is omitted certainly fits better. Palm Beach County now stands out, not as a Buchanan stronghold, but rather as a clear violation of the model that begs for explanation.

One of the great values of models is that, by establishing an idealized behavior, they help us to see when and how data values are unusual. In regression, a point can stand out in two different ways. First, a data value can have a large residual, as Palm Beach County does in this example. Because they seem to be different from the other cases, points whose residuals are large always deserve special attention.

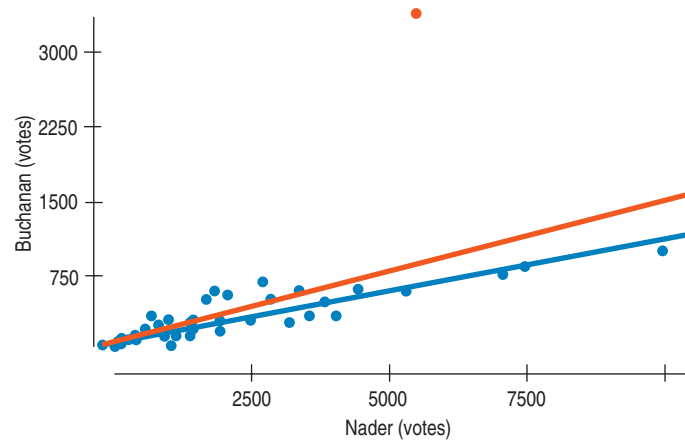


FIGURE 9.10

The red line shows the effect that one unusual point can have on a regression.



“Give me a place to stand and I will move the Earth.”

—Archimedes (287–211 BCE)

A data point can also be unusual if its  $x$ -value is far from the mean of the  $x$ -values. Such a point is said to have high **leverage**. The physical image of a lever is exactly right. We know the line must pass through  $(\bar{x}, \bar{y})$ , so you can picture that point as the fulcrum of the lever. Just as sitting farther from the hinge on a see-saw gives you more leverage to pull it your way, points with values far from  $\bar{x}$  pull more strongly on the regression line.

A point with high leverage has the potential to change the regression line. But it doesn’t always use that potential. If the point lines up with the pattern of the other points, then including it doesn’t change our estimate of the line. By sitting so far from  $\bar{x}$ , though, it may strengthen the relationship, inflating the correlation and  $R^2$ . How can you tell if a high-leverage point actually changes the model? Just fit the linear model twice, both with and without the point in question. We say that a point is **influential** if omitting it from the analysis gives a very different model.<sup>2</sup>

Influence depends on both leverage and residual; a case with high leverage whose  $y$ -value sits right on the line fit to the rest of the data is not influential.

**A S** **Activity: Leverage.** You may be surprised to see how sensitive to a single influential point a regression line is.

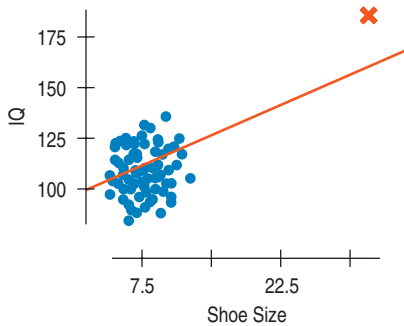
<sup>2</sup> Some textbooks use the term *influential point* for any observation that influences the slope, intercept, or  $R^2$ . We’ll reserve the term for points that influence the slope.

TI-*nspire*

**Influential points.** Try to make the regression line's slope change dramatically by dragging a point around in the scatterplot.

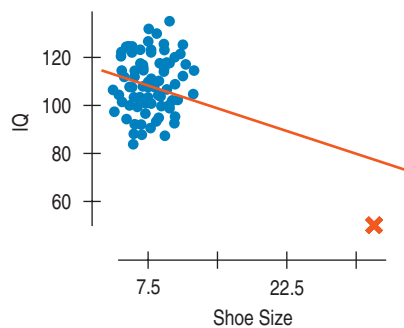
*“For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways.”*

—Francis Bacon  
(1561–1626)



**FIGURE 9.11**

Bozo's extraordinarily large shoes give his data point high leverage in the regression. Wherever Bozo's IQ falls, the regression line will follow.



**FIGURE 9.12**

If Bozo's IQ were low, the regression slope would change from positive to negative. A single influential point can change a regression model drastically.

Removing that case won't change the slope, even if it does affect  $R^2$ . A case with modest leverage but a very large residual (such as Palm Beach County) can be influential. Of course, if a point has enough leverage, it can pull the line right to it. Then it's highly influential, but its residual is small. The only way to be sure is to fit both regressions.

Unusual points in a regression often tell us more about the data and the model than any other points. We face a challenge: The best way to identify unusual points is against the background of a model, but good models are free of the influence of unusual points. (That insight's at least 400 years old. See the sidebar.) Don't give in to the temptation to simply delete points that don't fit the line. You can take points out and discuss what the model looks like with and without them, but arbitrarily deleting points can give a false sense of how well the model fits the data. Your goal should be understanding the data, not making  $R^2$  as big as you can.

In 2000, George W. Bush won Florida (and thus the presidency) by only a few hundred votes, so Palm Beach County's residual is big enough to be meaningful. It's the rare unusual point that determines a presidency, but all are worth examining and trying to understand.

A point with so much influence that it pulls the regression line close to it can make its residual deceptively small. Influential points like that can have a shocking effect on the regression. Here's a plot of IQ against Shoe Size, again from the fanciful study of intelligence and foot size in comedians we saw in Chapter 7. The linear regression output shows

Dependent variable is: IQ

R-squared = 24.8%

Variable	Coefficient
Intercept	93.3265
Shoe size	2.08318

Although this is a silly example, it illustrates an important and common potential problem: Almost all of the variance accounted for ( $R^2 = 24.8\%$ ) is due to one point, namely, Bozo. Without Bozo, there is little correlation between Shoe Size and IQ. Look what happens to the regression when we take him out:

Dependent variable is: IQ

R-squared = 0.7%

Variable	Coefficient
Intercept	105.458
Shoe size	-0.460194

The  $R^2$  value is now 0.7%—a very weak linear relationship (as one might expect!). One single point exhibits a great influence on the regression analysis.

What would have happened if Bozo hadn't shown his comic genius on IQ tests? Suppose his measured IQ had been only 50. The slope of the line would then drop from 0.96 IQ points/shoe size to  $-0.69$  IQ points/shoe size. No matter where Bozo's IQ is, the line tends to follow it because his Shoe Size, being so far from the mean Shoe Size, makes this a high-leverage point.

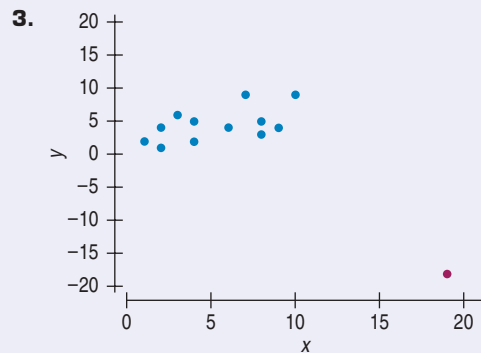
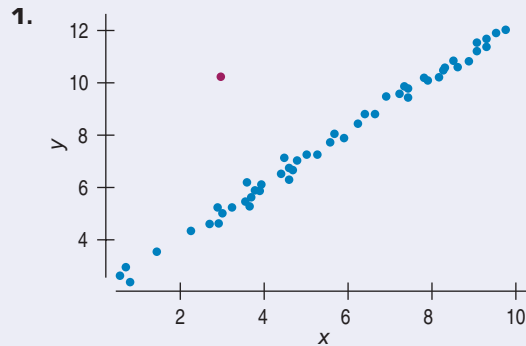
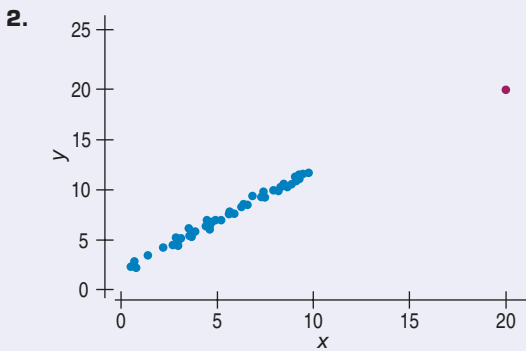
Even though this example is far fetched, similar situations occur all the time in real life. For example, a regression of sales against floor space for hardware stores that looked primarily at small-town businesses could be dominated in a similar way if The Home Depot were included.

**Warning:** Influential points can hide in plots of residuals. Points with high leverage pull the line close to them, so they often have small residuals. You'll see influential points more easily in scatterplots of the original data or by finding a regression model with and without the points.



### JUST CHECKING

Each of these scatterplots shows an unusual point. For each, tell whether the point is a high-leverage point, would have a large residual, or is influential.

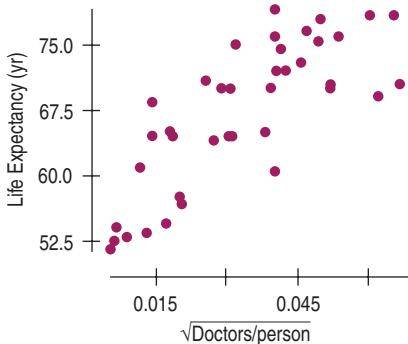


## Lurking Variables and Causation

One common way to interpret a regression slope is to say that “a change of 1 unit in  $x$  results in a change of  $b_1$  units in  $y$ .” This way of saying things encourages causal thinking. Beware.

In Chapter 7, we tried to make it clear that no matter how strong the correlation is between two variables, there’s no simple way to show that one variable causes the other. Putting a regression line through a cloud of points just increases the temptation to think and to say that the  $x$ -variable *causes* the  $y$ -variable. Just to make sure, let’s repeat the point again: **No matter how strong the association, no matter how large the  $R^2$  value, no matter how straight the line, there is no way to conclude from a regression alone that one variable *causes* the other.** There’s always the possibility that some third variable is driving both of the variables you have observed. **With observational data, as opposed to data from a designed experiment, there is no way to be sure that a **lurking variable** is not the cause of any apparent association.**

Here’s an example: The scatterplot shows the *Life Expectancy* (average of men and women, in years) for each of 41 countries of the world, plotted against the square root of the number of *Doctors* per person in the country. (The square root is here to make the relationship satisfy the Straight Enough Condition, as we saw back in Chapter 7.)



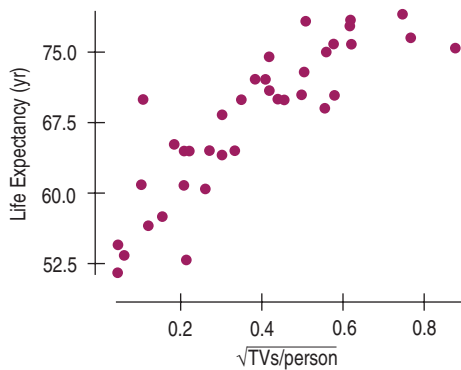
**FIGURE 9.13**

The relationship between Life Expectancy (years) and availability of Doctors (measured as  $\sqrt{\text{doctors per person}}$ ) for countries of the world is strong, positive, and linear.

The strong positive association ( $R^2 = 62.4\%$ ) seems to confirm our expectation that more *Doctors* per person improves healthcare, leading to longer lifetimes and a greater *Life Expectancy*. The strength of the association would *seem* to argue that we should send more doctors to developing countries to increase life expectancy.

That conclusion is about the consequences of a change. Would sending more doctors increase life expectancy? Specifically, do doctors *cause* greater life expectancy? Perhaps, but these are observed data, so there may be another explanation for the association.

On the next page, the similar-looking scatterplot’s  $x$ -variable is the square root of the number of *Televisions* per person in each country. The positive association in this scatterplot is even *stronger* than the association in the previous plot



**FIGURE 9.14**

To increase life expectancy, don't send doctors, send TVs; they're cheaper and more fun. Or maybe that's not the right interpretation of this scatterplot of life expectancy against availability of TVs (as  $\sqrt{\text{TVs per person}}$ ).

( $R^2 = 72.3\%$ ). We can fit the linear model, and quite possibly use the number of TVs as a way to predict life expectancy. Should we conclude that increasing the number of TVs actually extends lifetimes? If so, we should send TVs instead of doctors to developing countries. Not only is the correlation with life expectancy higher, but TVs are much cheaper than doctors.

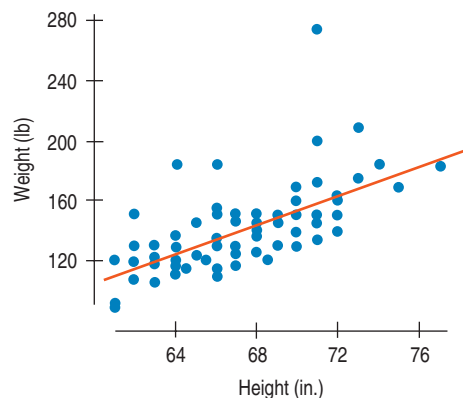
What's wrong with this reasoning? Maybe we were a bit hasty earlier when we concluded that doctors *cause* longer lives. Maybe there's a lurking variable here. Countries with higher standards of living have both longer life expectancies *and* more doctors (and more TVs). Could higher living standards cause changes in the other variables? If so, then improving living standards might be expected to prolong lives, increase the number of doctors, and increase the number of TVs.

From this example, you can see how easy it is to fall into the trap of mistakenly inferring causality from a regression. For all we know, doctors (or TVs!) *do* increase life expectancy. But we can't tell that from data like these, no matter how much we'd like to. Resist the temptation to conclude that  $x$  causes  $y$  from a regression, no matter how obvious that conclusion seems to you.

## Working with Summary Values

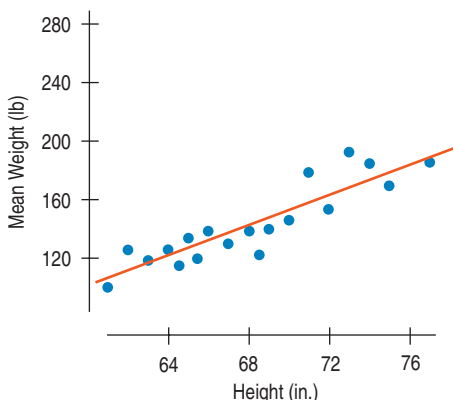
Scatterplots of statistics summarized over groups tend to show less variability than we would see if we measured the same variable on individuals. This is because the summary statistics themselves vary less than the data on the individuals do—a fact we will make more specific in coming chapters.

In Chapter 7 we looked at the heights and weights of individual students. There we saw a correlation of 0.644, so  $R^2$  is 41.5%.



**FIGURE 9.15**

Weight (lb) against Height (in.) for a sample of men. There's a strong, positive, linear association.



**FIGURE 9.16**

Mean Weight (lb) shows a stronger linear association with Height than do the weights of individuals. Means vary less than individual values.

Suppose, instead of data on individuals, we knew only the mean weight for each height value. The scatterplot of mean weight by height would show less scatter. And the  $R^2$  would increase to 80.1%.

Scatterplots of summary statistics show less scatter than the baseline data on individuals and can give a false impression of how well a line summarizes the data. There's no simple correction for this phenomenon. Once we're given summary data, there's no simple way to get the original values back.

In the life expectancy and TVs example, we have no good measure of exposure to doctors or to TV on an individual basis. But if we did, we should expect the scatterplot to show more variability and the corresponding  $R^2$  to be smaller. The bottom line is that you should be a bit suspicious of conclusions based on regressions of summary data. They may look better than they really are.

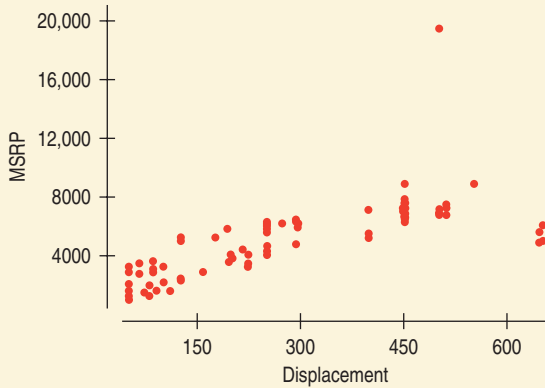
**FOR EXAMPLE**

Using several of these methods together

Motorcycles designed to run off-road, often known as dirt bikes, are specialized vehicles.

We have data on 104 dirt bikes available for sale in 2005. Some cost as little as \$3000, while others are substantially more expensive. Let's investigate how the size and type of engine contribute to the cost of a dirt bike. As always, we start with a scatterplot.

Here's a scatterplot of the manufacturer's suggested retail price (*MSRP*) in dollars against the engine *Displacement*, along with a regression analysis:



Dependent variable is: MSRP

R-squared = 49.9%  $s_e = 1737$

Variable	Coefficient
Intercept	2273.67
Displacement	10.0297

**Question:** What do you see in the scatterplot?

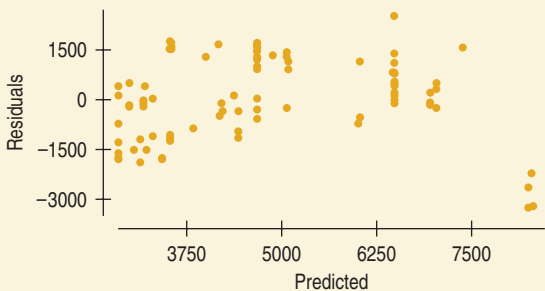
There is a strong positive association between the engine displacement of dirt bikes and the manufacturer's suggested retail price. One of the dirt bikes is an outlier; its price is more than double that of any other bike.

The outlier is the Husqvarna TE 510 Centennial. Most of its components are handmade exclusively for this model, including extensive use of carbon fiber throughout. That may explain its \$19,500 price tag! Clearly, the TE 510 is not like the other bikes. We'll set it aside for now and look at the data for the remaining dirt bikes.

**Question:** What effect will removing this outlier have on the regression? Describe how the slope,  $R^2$ , and  $s_e$  will change.

The TE 510 was an influential point, tilting the regression line upward. With that point removed, the regression slope will get smaller. With that dirt bike omitted, the pattern becomes more consistent, so the value of  $R^2$  should get larger and the standard deviation of the residuals,  $s_e$ , should get smaller.

With the outlier omitted, here's the new regression and a scatterplot of the residuals:



Dependent variable is: MSRP

R-squared = 61.3%  $s_e = 1237$

Variable	Coefficient
Intercept	2411.02
Displacement	9.05450

**Question:** What do you see in the residuals plot?

The points at the far right don't fit well with the other dirt bikes. Overall, there appears to be a bend in the relationship, so a linear model may not be appropriate.

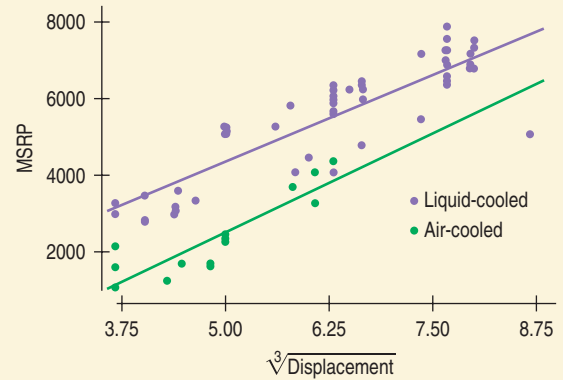
Let's try a re-expression. Here's a scatterplot showing *MSRP* against the cube root of *Displacement* to make the relationship closer to straight. (Since displacement is measured in cubic centimeters, its cube root has the simple units of centimeters.) In addition, we've colored the plot according to the cooling

method used in the bike's engine: liquid or air. Each group is shown with its own regression line, as we did for the cereals on different shelves.

**Question:** What does this plot say about dirt bikes?

There appears to be a positive, linear relationship between MSRP and the cube root of Displacement. In general, the larger the engine a bike has, the higher the suggested price. Liquid-cooled dirt bikes, however, typically cost more than air-cooled bikes with comparable displacement. A few liquid-cooled bikes appear to be much less expensive than we might expect, given their engine displacements.

[Jiang Lu, Joseph B. Kadane, and Peter Boatwright, "The Dirt on Bikes: An Illustration of CART Models for Brand Differentiation," provides data on 2005-model bikes.]



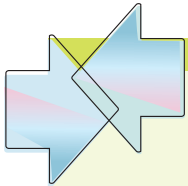
## WHAT CAN GO WRONG?

This entire chapter has held warnings about things that can go wrong in a regression analysis. So let's just recap. When you make a linear model:

- ▶ **Make sure the relationship is straight.** Check the Straight Enough Condition. Always examine the residuals for evidence that the Linearity Assumption has failed. It's often easier to see deviations from a straight line in the residuals plot than in the scatterplot of the original data. Pay special attention to the most extreme residuals because they may have something to add to the story told by the linear model.
- ▶ **Be on guard for different groups in your regression.** Check for evidence that the data consist of separate subsets. If you find subsets that behave differently, consider fitting a different linear model to each subset.
- ▶ **Beware of extrapolating.** Beware of extrapolation beyond the  $x$ -values that were used to fit the model. Although it's common to use linear models to extrapolate, the practice is dangerous.
- ▶ **Beware especially of extrapolating into the future!** Be especially cautious about extrapolating into the future with linear models. To predict the future, you must assume that future changes will continue at the same rate you've observed in the past. Predicting the future is particularly tempting and particularly dangerous.
- ▶ **Look for unusual points.** Unusual points always deserve attention and may well reveal more about your data than the rest of the points combined. Always look for them and try to understand why they stand apart. A scatterplot of the data is a good way to see high-leverage and influential points. A scatterplot of the residuals against the predicted values is a good tool for finding points with large residuals.
- ▶ **Beware of high-leverage points and especially of those that are influential.** Influential points can alter the regression model a great deal. The resulting model may say more about one or two points than about the overall relationship.
- ▶ **Consider comparing two regressions.** To see the impact of outliers on a regression, it's often wise to run two regressions, one with and one without the extraordinary points, and then to discuss the differences.
- ▶ **Treat unusual points honestly.** If you remove enough carefully selected points, you can always get a regression with a high  $R^2$  eventually. But it won't give you much understanding. Some variables are not related in a way that's simple enough for a linear model to fit very well. When that happens, report the failure and stop.

(continued)

- ▶ **Beware of lurking variables.** Think about lurking variables before interpreting a linear model. It's particularly tempting to explain a strong regression by thinking that the  $x$ -variable *causes* the  $y$ -variable. A linear model alone can never demonstrate such causation, in part because it cannot eliminate the chance that a lurking variable has caused the variation in both  $x$  and  $y$ .
- ▶ **Watch out when dealing with data that are summaries.** Be cautious in working with data values that are themselves summaries, such as means or medians. Such statistics are less variable than the data on which they are based, so they tend to inflate the impression of the strength of a relationship.



## CONNECTIONS

We are always alert to things that can go wrong if we use statistics without thinking carefully. Regression opens new vistas of potential problems. But each one relates to issues we've thought about before.

It is always important that our data be from a single homogeneous group and not made up of disparate groups. We looked for multiple modes in single variables. Now we check scatterplots for evidence of subgroups in our data. As with modes, it's often best to split the data and analyze the groups separately.

Our concern with unusual points and their potential influence also harks back to our earlier concern with outliers in histograms and boxplots—and for many of the same reasons. As we've seen here, regression offers such points new scope for mischief.

The risks of interpreting linear models as causal or predictive arose in Chapters 7 and 8. And they're important enough to mention again in later chapters.



## WHAT HAVE WE LEARNED?

We've learned that there are many ways in which a data set may be unsuitable for a regression analysis.

- ▶ Watch out for more than one group hiding in your regression analysis. If you find subsets of the data that behave differently, consider fitting a different regression model to each subset.
- ▶ The **Straight Enough Condition** says that the relationship should be reasonably straight to fit a regression. Somewhat paradoxically, sometimes it's easier to see that the relationship is not straight *after* fitting the regression by examining the residuals. The same is true of outliers.
- ▶ The **Outlier Condition** actually means two things: Points with large residuals or high leverage (especially both) can influence the regression model significantly. It's a good idea to perform the regression analysis with and without such points to see their impact.

And we've learned that even a good regression doesn't mean we should believe that the model says more than it really does.

- ▶ Extrapolation far from  $\bar{x}$  can lead to silly and useless predictions.
- ▶ Even an  $R^2$  near 100% doesn't indicate that  $x$  causes  $y$  (or the other way around). Watch out for lurking variables that may affect both  $x$  and  $y$ .
- ▶ Be careful when you interpret regressions based on *summaries* of the data sets. These regressions tend to look stronger than the regression based on all the individual data.

## Terms

### Extrapolation

203. Although linear models provide an easy way to predict values of  $y$  for a given value of  $x$ , it is unsafe to predict for values of  $x$  far from the ones used to find the linear model equation. Such extrapolation may pretend to see into the future, but the predictions should not be trusted.



Outlier	205. Any data point that stands away from the others can be called an outlier. In regression, outliers can be extraordinary in two ways: by having a large residual or by having high leverage.
Leverage	206. Data points whose $x$ -values are far from the mean of $x$ are said to exert leverage on a linear model. High-leverage points pull the line close to them, and so they can have a large effect on the line, sometimes completely determining the slope and intercept. With high enough leverage, their residuals can be deceptively small.
Influential point	206. If omitting a point from the data results in a very different regression model, then that point is called an influential point.
Lurking variable	208. A variable that is not explicitly part of a model but affects the way the variables in the model appear to be related is called a lurking variable. Because we can never be certain that observational data are not hiding a lurking variable that influences both $x$ and $y$ , it is never safe to conclude that a linear model demonstrates a causal relationship, no matter how strong the linear association.

## Skills

### THINK

- ▶ Understand that we cannot fit linear models or use linear regression if the underlying relationship between the variables is not itself linear.
- ▶ Understand that data used to find a model must be homogeneous. Look for subgroups in data before you find a regression, and analyze each separately.
- ▶ Know the danger of extrapolating beyond the range of the  $x$ -values used to find the linear model, especially when the extrapolation tries to predict into the future.
- ▶ Understand that points can be unusual by having a large residual or by having high leverage.
- ▶ Understand that an influential point can change the slope and intercept of the regression line.
- ▶ Look for lurking variables whenever you consider the association between two variables. Understand that a strong association does not mean that the variables are causally related.
- ▶ Know how to display residuals from a linear model by making a scatterplot of residuals against predicted values or against the  $x$ -variable, and know what patterns to look for in the picture.

### SHOW

- ▶ Know how to look for high-leverage and influential points by examining a scatterplot of the data and how to look for points with large residuals by examining a scatterplot of the residuals against the predicted values or against the  $x$ -variable. Understand how fitting a regression line with and without influential points can add to your understanding of the regression model.
- ▶ Know how to look for high-leverage points by examining the distribution of the  $x$ -values or by recognizing them in a scatterplot of the data, and understand how they can affect a linear model.

### TELL

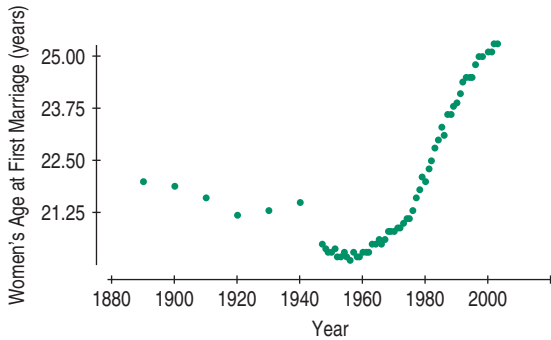
- ▶ Include diagnostic information such as plots of residuals and leverages as part of your report of a regression.
- ▶ Report any high-leverage points.
- ▶ Report any outliers. Consider reporting analyses with and without outliers, to assess their influence on the regression.
- ▶ Include appropriate cautions about extrapolation when reporting predictions from a linear model.
- ▶ Discuss possible lurking variables.

## REGRESSION DIAGNOSIS ON THE COMPUTER

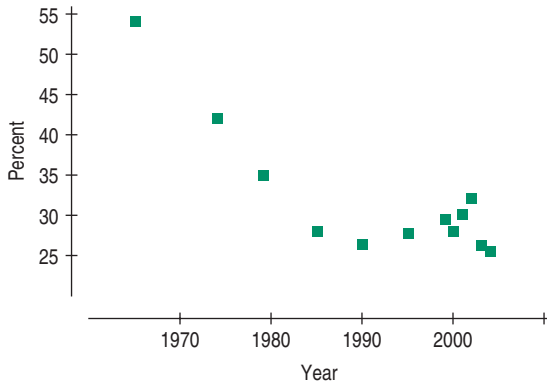
Most statistics technology offers simple ways to check whether your data satisfy the conditions for regression. We have already seen that these programs can make a simple scatterplot. They can also help us check the conditions by plotting residuals.

## EXERCISES

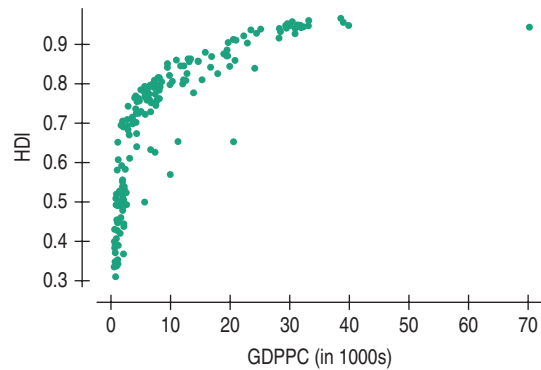
- T** 1. **Marriage age 2003.** Is there evidence that the age at which women get married has changed over the past 100 years? The scatterplot shows the trend in age at first marriage for American women ([www.census.gov](http://www.census.gov)).



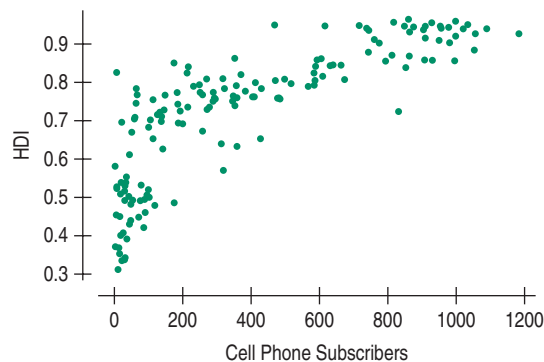
- a) Is there a clear pattern? Describe the trend.  
 b) Is the association strong?  
 c) Is the correlation high? Explain.  
 d) Is a linear model appropriate? Explain.
- T** 2. **Smoking 2004.** The Centers for Disease Control and Prevention track cigarette smoking in the United States. How has the percentage of people who smoke changed since the danger became clear during the last half of the 20th century? The scatterplot shows percentages of smokers among men 18–24 years of age, as estimated by surveys, from 1965 through 2004 ([www.cdc.gov/nchs/](http://www.cdc.gov/nchs/)).



- a) Is there a clear pattern? Describe the trend.  
 b) Is the association strong?  
 c) Is a linear model appropriate? Explain.
- T** 3. **Human Development Index.** The United Nations Development Programme (UNDP) uses the Human Development Index (HDI) in an attempt to summarize in one number the progress in health, education, and economics of a country. In 2006, the HDI was as high as 0.965 for Norway and as low as 0.331 for Niger. The gross domestic product per capita (GDPPC), by contrast, is often used to summarize the *overall* economic strength of a country. Is the HDI related to the GDPPC? Here is a scatterplot of HDI against GDPPC.



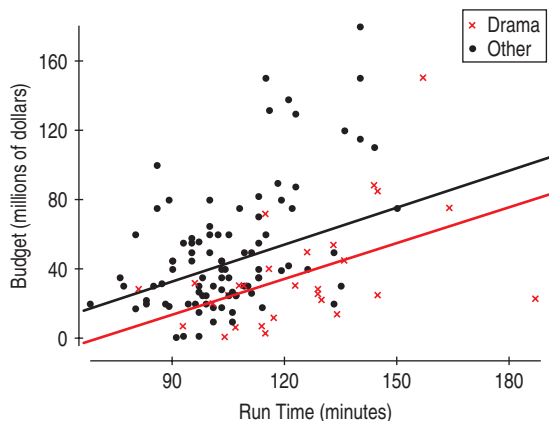
- a) Explain why fitting a linear model to these data might be misleading.  
 b) If you fit a linear model to the data, what do you think a scatterplot of residuals versus predicted HDI will look like?  
 c) There is an outlier (Luxembourg) with a GDPPC of around \$70,000. Will setting this point aside improve the model substantially? Explain.
- T** 4. **HDI Revisited.** The United Nations Development Programme (UNDP) uses the Human Development Index (HDI) in an attempt to summarize in one number the progress in health, education, and economics of a country. The number of cell phone subscribers per 1000 people is positively associated with economic progress in a country. Can the number of cell phone subscribers be used to predict the HDI? Here is a scatterplot of HDI against cell phone subscribers:



- a) Explain why fitting a linear model to these data might be misleading.  
 b) If you fit a linear model to the data, what do you think a scatterplot of residuals versus predicted HDI will look like?
5. **Good model?** In justifying his choice of a model, a student wrote, "I know this is the correct model because  $R^2 = 99.4\%$ ."  
 a) Is this reasoning correct? Explain.  
 b) Does this model allow the student to make accurate predictions? Explain.

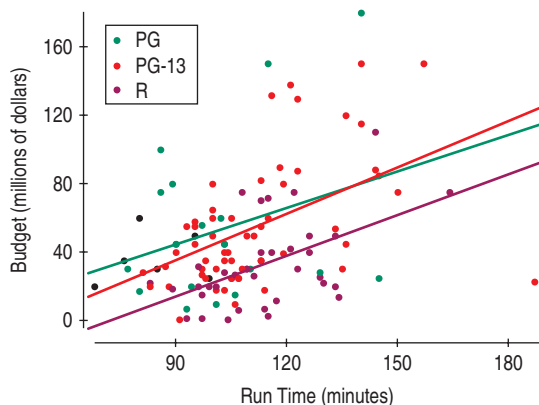
6. **Bad model?** A student who has created a linear model is disappointed to find that her  $R^2$  value is a very low 13%.
- Does this mean that a linear model is not appropriate? Explain.
  - Does this model allow the student to make accurate predictions? Explain.

**T** 7. **Movie Dramas.** Here's a scatterplot of the production budgets (in millions of dollars) vs. the running time (in minutes) for major release movies in 2005. Dramas are plotted in red and all other genres are plotted in black. A separate least squares regression line has been fitted to each group. For the following questions, just examine the plot:



- What are the units for the slopes of these lines?
- In what way are dramas and other movies similar with respect to this relationship?
- In what way are dramas different from other genres of movies with respect to this relationship?

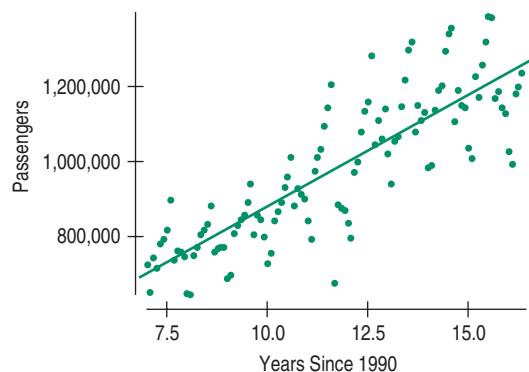
**T** 8. **Movie Ratings.** Does the cost of making a movie depend on its audience? Here's a scatterplot of the same data we examined in Exercise 7. Movies with an R rating are colored purple, those with a PG-13 rating are red, and those with a PG rating are green. Regression lines have been found for each group. (The black points are G-rated, but there were too few to fit a line reliably.)



- In what ways is the relationship between run times and budgets similar for the three ratings groups?
- How do the costs of R-rated movies differ from those of PG-13 and PG rated movies? Discuss both the slopes and the intercepts.

- The film *King Kong*, with a run time of 187 minutes, is the red point sitting at the lower right. If it were omitted from this analysis, how might that change your conclusions about PG-13 movies?

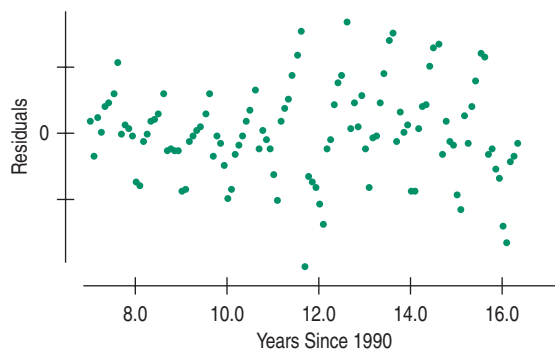
**T** 9. **Oakland passengers.** The scatterplot below shows the number of passengers departing from Oakland (CA) airport month by month since the start of 1997. Time is shown as years since 1990, with fractional years used to represent each month. (Thus, June of 1997 is 7.5—halfway through the 7th year after 1990.) [www.oaklandairport.com](http://www.oaklandairport.com)



Here's a regression and the residuals plot:

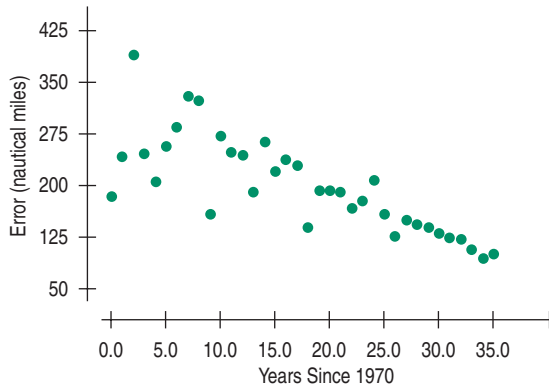
Dependent variable is: Passengers  
 $R$ -squared = 71.1 %  $s_e$  = 104330

Variable	Coefficient
Constant	282584
Year-1990	59704.4



- Interpret the slope and intercept of the model.
- What does the value of  $R^2$  say about the model?
- Interpret  $s_e$  in this context.
- Would you use this model to predict the numbers of passengers in 2010 ( $YearsSince1990 = 20$ )? Explain.
- There's a point near the middle of this time span with a large negative residual. Can you explain this outlier?

**T** 10. **Tracking hurricanes.** In a previous chapter, we saw data on the errors (in nautical miles) made by the National Hurricane Center in predicting the path of hurricanes. The scatterplot on the next page shows the trend in the 24-hour tracking errors since 1970 ([www.nhc.noaa.gov](http://www.nhc.noaa.gov)).

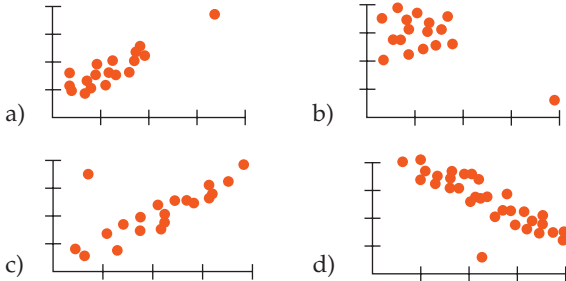


Dependent variable is: Error  
 R-squared = 63.0 %  $s_e = 42.87$

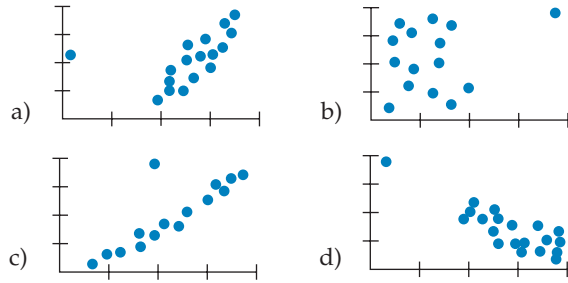
Variable	Coefficient
Intercept	292.089
Years-1970	-5.22924

- Interpret the slope and intercept of the model.
- Interpret  $s_e$  in this context.
- The Center had a stated goal of achieving an average tracking error of 125 nautical miles in 2009. Will they make it? Why do you think so?
- What if their goal were an average tracking error of 90 nautical miles?
- What cautions would you state about your conclusion?

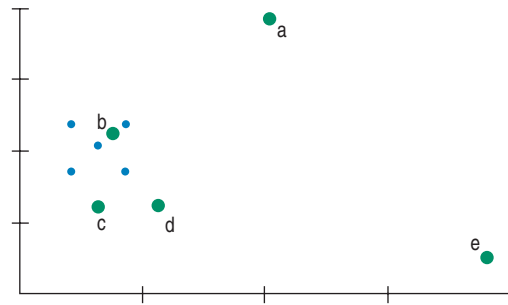
11. **Unusual points.** Each of the four scatterplots that follow shows a cluster of points and one “stray” point. For each, answer these questions:
- In what way is the point unusual? Does it have high leverage, a large residual, or both?
  - Do you think that point is an influential point?
  - If that point were removed, would the correlation become stronger or weaker? Explain.
  - If that point were removed, would the slope of the regression line increase or decrease? Explain.



12. **More unusual points.** Each of the following scatterplots shows a cluster of points and one “stray” point. For each, answer these questions:
- In what way is the point unusual? Does it have high leverage, a large residual, or both?
  - Do you think that point is an influential point?
  - If that point were removed, would the correlation become stronger or weaker? Explain.
  - If that point were removed, would the slope of the regression line increase or decrease? Explain.



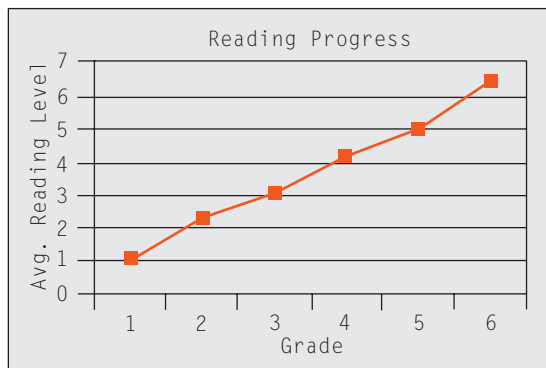
13. **The extra point.** The scatterplot shows five blue data points at the left. Not surprisingly, the correlation for these points is  $r = 0$ . Suppose *one* additional data point is added at one of the five positions suggested below in green. Match each point (a–e) with the correct new correlation from the list given.
- |            |           |
|------------|-----------|
| 1) $-0.90$ | 4) $0.05$ |
| 2) $-0.40$ | 5) $0.75$ |
| 3) $0.00$  |           |



14. **The extra point revisited.** The original five points in Exercise 13 produce a regression line with slope 0. Match each of the green points (a–e) with the slope of the line after that one point is added:
- |            |           |
|------------|-----------|
| 1) $-0.45$ | 4) $0.05$ |
| 2) $-0.30$ | 5) $0.85$ |
| 3) $0.00$  |           |

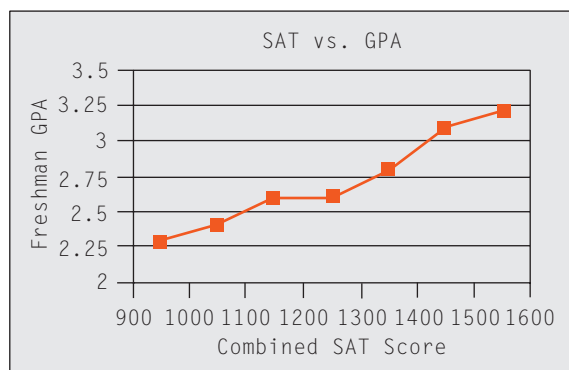
15. **What’s the cause?** Suppose a researcher studying health issues measures blood pressure and the percentage of body fat for several adult males and finds a strong positive association. Describe three different possible cause-and-effect relationships that might be present.
16. **What’s the effect?** A researcher studying violent behavior in elementary school children asks the children’s parents how much time each child spends playing computer games and has their teachers rate each child on the level of aggressiveness they display while playing with other children. Suppose that the researcher finds a moderately strong positive correlation. Describe three different possible cause-and-effect explanations for this relationship.
17. **Reading.** To measure progress in reading ability, students at an elementary school take a reading comprehension test every year. Scores are measured in “grade-level” units; that is, a score of 4.2 means that a student is reading at slightly above the expected level for a fourth grader. The school principal prepares a report to parents that includes a graph showing the mean reading score for

each grade. In his comments he points out that the strong positive trend demonstrates the success of the school's reading program.

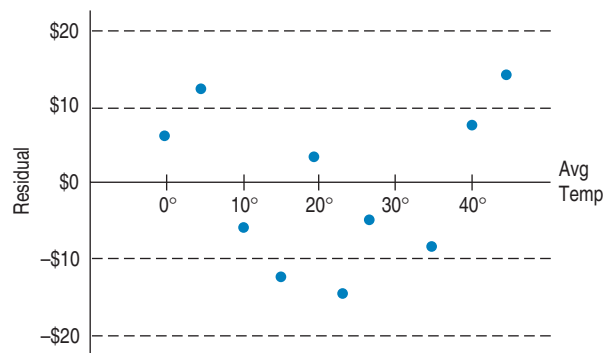


- Does this graph indicate that students are making satisfactory progress in reading? Explain.
- What would you estimate the correlation between *Grade* and *Average Reading Level* to be?
- If, instead of this plot showing average reading levels, the principal had produced a scatterplot of the reading levels of all the individual students, would you expect the correlation to be the same, higher, or lower? Explain.
- Although the principal did not do a regression analysis, someone as statistically astute as you might do that. (But don't bother.) What value of the slope of that line would you view as demonstrating acceptable progress in reading comprehension? Explain.

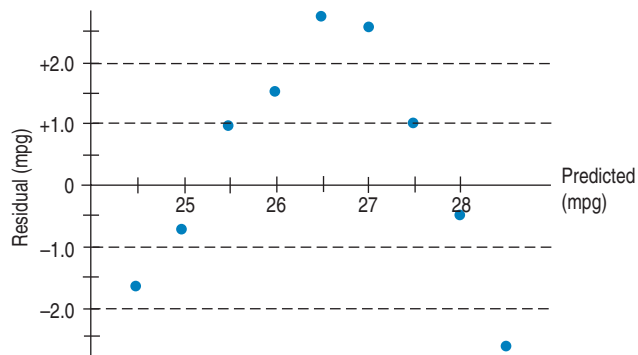
18. **Grades.** A college admissions officer, defending the college's use of SAT scores in the admissions process, produced the graph below. It shows the mean GPAs for last year's freshmen, grouped by SAT scores. How strong is the evidence that *SAT Score* is a good predictor of *GPA*? What concerns you about the graph, the statistical methodology or the conclusions reached?



19. **Heating.** After keeping track of his heating expenses for several winters, a homeowner believes he can estimate the monthly cost from the average daily Fahrenheit temperature by using the model  $\widehat{Cost} = 133 - 2.13 Temp$ . Here is the residuals plot for his data:

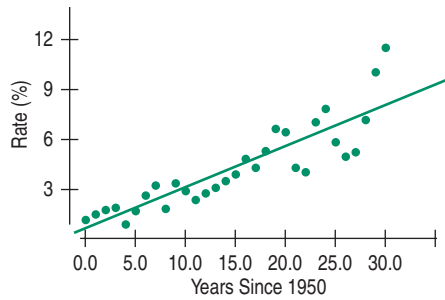


- Interpret the slope of the line in this context.
  - Interpret the  $y$ -intercept of the line in this context.
  - During months when the temperature stays around freezing, would you expect cost predictions based on this model to be accurate, too low, or too high? Explain.
  - What heating cost does the model predict for a month that averages  $10^\circ$ ?
  - During one of the months on which the model was based, the temperature did average  $10^\circ$ . What were the actual heating costs for that month?
  - Should the homeowner use this model? Explain.
  - Would this model be more successful if the temperature were expressed in degrees Celsius? Explain.
20. **Speed.** How does the speed at which you drive affect your fuel economy? To find out, researchers drove a compact car for 200 miles at speeds ranging from 35 to 75 miles per hour. From their data, they created the model  $\widehat{Fuel\ Efficiency} = 32 - 0.1 Speed$  and created this residual plot:



- Interpret the slope of this line in context.
- Explain why it's silly to attach any meaning to the  $y$ -intercept.
- When this model predicts high *Fuel Efficiency*, what can you say about those predictions?
- What *Fuel Efficiency* does the model predict when the car is driven at 50 mph?
- What was the actual *Fuel Efficiency* when the car was driven at 45 mph?
- Do you think there appears to be a strong association between *Speed* and *Fuel Efficiency*? Explain.
- Do you think this is the appropriate model for that association? Explain.

**T 21. Interest rates.** Here's a plot showing the federal rate on 3-month Treasury bills from 1950 to 1980, and a regression model fit to the relationship between the *Rate* (in %) and *Years since 1950* ([www.gpoaccess.gov/eop/](http://www.gpoaccess.gov/eop/)).

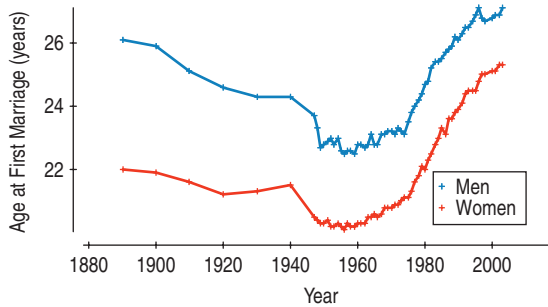


Dependent variable is: Rate  
R-squared = 77.4 % s = 1.239

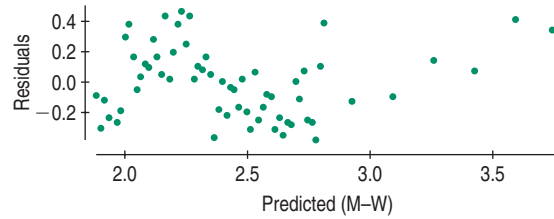
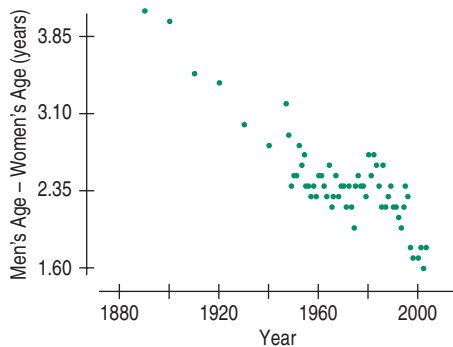
Variable	Coefficient
Intercept	0.640282
Year - 1950	0.247637

- What is the correlation between *Rate* and *Year*?
- Interpret the slope and intercept.
- What does this model predict for the interest rate in the year 2000?
- Would you expect this prediction to have been accurate? Explain.

**22. Ages of couples 2003.** The graph shows the ages of both men and women at first marriage ([www.census.gov](http://www.census.gov)).



Clearly, the pattern for men is similar to the pattern for women. But are the two lines getting closer together? Here's a timeplot showing the *difference* in average age (men's age - women's age) at first marriage, the regression analysis, and the associated residuals plot.

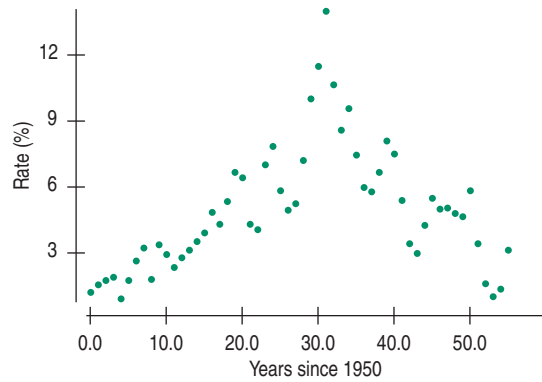


Dependent variable is: Age Difference  
R-squared = 75.1 % s = 0.2333

Variable	Coefficient
Constant	35.0617
Year	-0.016565

- What is the correlation between *Age Difference* and *Year*?
- Interpret the slope of this line.
- Predict the average age difference in 2015.
- Describe reasons why you might not place much faith in that prediction.

**T 23. Interest rates revisited.** In Exercise 21 you investigated the federal rate on 3-month Treasury bills between 1950 and 1980. The scatterplot below shows that the trend changed dramatically after 1980.



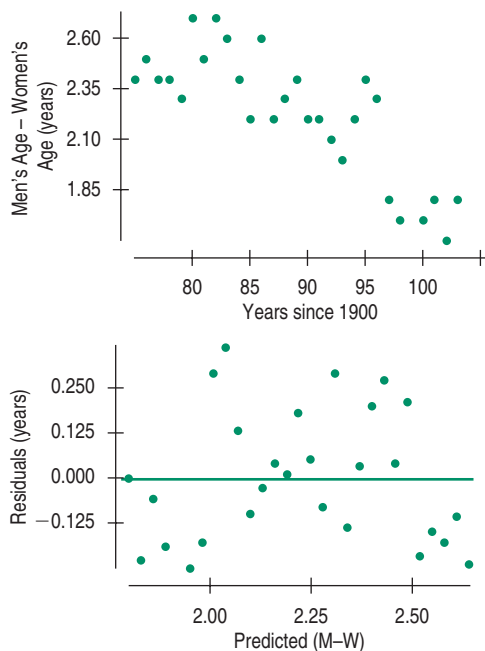
Here's a regression model for the data since 1980.

Dependent variable is: Rate  
R-squared = 74.5 % s = 1.630

Variable	Coefficient
Intercept	21.0688
Year - 1950	-0.356578

- How does this model compare to the one in Exercise 21?
- What does this model estimate the interest rate to have been in 2000? How does this compare to the rate you predicted in Exercise 21?
- Do you trust this newer predicted value? Explain.
- Given these two models, what would you predict the interest rate on 3-month Treasury bills will be in 2020?

**T 24. Ages of couples, again.** Has the trend of decreasing difference in age at first marriage seen in Exercise 22 gotten stronger recently? The scatterplot and residual plot for the data from 1975 through 2003, along with a regression for just those years, are on the next page.

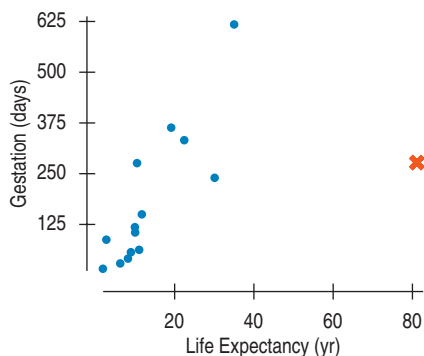


Dependent variable is: Men - Women  
 R-Squared = 65.6 % s = 0.1869

Variable	Coefficient
Intercept	4.88424
Year	-0.029959

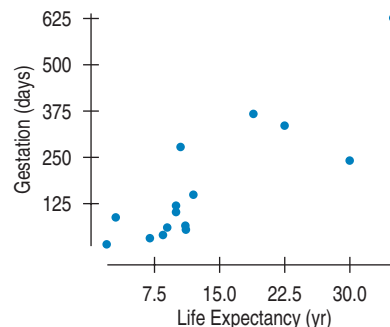
- Why is  $R^2$  higher for the first model (in Exercise 22)?
- Is this linear model appropriate for the post-1975 data? Explain.
- What does the slope say about marriage ages?
- Explain why it's not reasonable to interpret the  $y$ -intercept.

25. **Gestation.** For women, pregnancy lasts about 9 months. In other species of animals, the length of time from conception to birth varies. Is there any evidence that the gestation period is related to the animal's lifespan? The first scatterplot shows *Gestation Period* (in days) vs. *Life Expectancy* (in years) for 18 species of mammals. The highlighted point at the far right represents humans.



- For these data,  $r = 0.54$ , not a very strong relationship. Do you think the association would be stronger or weaker if humans were removed? Explain.
- Is there reasonable justification for removing humans from the data set? Explain.

c) Here are the scatterplot and regression analysis for the 17 nonhuman species. Comment on the strength of the association.



Dependent variable is: Gestation  
 R-Squared = 72.2 %

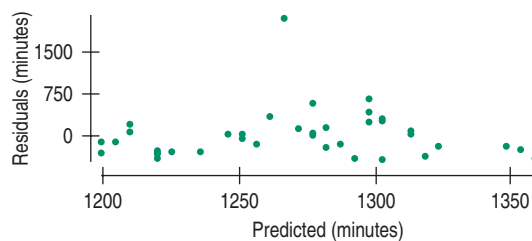
Variable	Coefficient
Constant	-39.5172
Lif Exp	15.4980

- Interpret the slope of the line.
- Some species of monkeys have a life expectancy of about 20 years. Estimate the expected gestation period of one of these monkeys.

T 26. **Swim the lake 2006.** People swam across Lake Ontario 42 times between 1974 and 2006 ([www.soloswims.com](http://www.soloswims.com)). We might be interested in whether they are getting any faster or slower. Here are the regression of the crossing *Times* (minutes) against the *Year* of the crossing and the residuals plot:

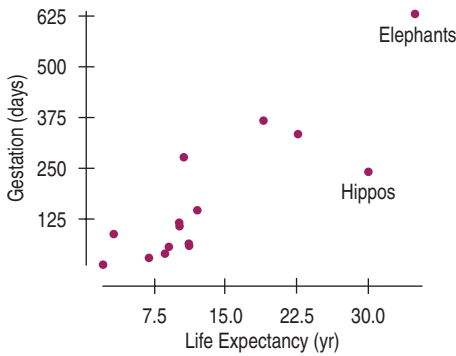
Dependent variable is: Time  
 R-Squared = 1.3 % s = 443.8

Variable	Coefficient
Intercept	-8950.40
Year	5.14171



- What does the  $R^2$  mean for this regression?
- Are the swimmers getting faster or slower? Explain.
- The outlier seen in the residuals plot is a crossing by Vicki Keith in 1987 in which she swam a round trip, north to south, and then back again. Clearly, this swim doesn't belong with the others. Would removing it change the model a lot? Explain.

27. **Elephants and hippos.** We removed humans from the scatterplot in Exercise 25 because our species was an outlier in life expectancy. The resulting scatterplot (next page) shows two points that now may be of concern. The point in the upper right corner of this scatterplot is for elephants, and the other point at the far right is for hippos.



- By removing one of these points, we could make the association appear to be stronger. Which point? Explain.
- Would the slope of the line increase or decrease?
- Should we just keep removing animals to increase the strength of the model? Explain.
- If we remove elephants from the scatterplot, the slope of the regression line becomes 11.6 days per year. Do you think elephants were an influential point? Explain.

**T 28. Another swim 2006.** In Exercise 26 we saw that Vicki Keith’s round-trip swim of Lake Ontario was an obvious outlier among the other one-way times. Here is the new regression after this unusual point is removed:

Dependent variable is: Time  
 R-Squared = 4.1 %    $s_e = 292.6$

Variable	Coefficient
Intercept	-11048.7
Year	6.17091

- In this new model, the value of  $s_e$  is much smaller. Explain what that means in this context.
- Now would you be willing to say that the Lake Ontario swimmers are getting faster (or slower)?

**T 29. Marriage age 2003 revisited.** Suppose you wanted to predict the trend in marriage age for American women into the early part of this century.

- How could you use the data graphed in Exercise 1 to get a good prediction? Marriage ages in selected years starting in 1900 are listed below. Use all or part of these data to create an appropriate model for predicting the average age at which women will first marry in 2010.

1900–1950 (10-yr intervals): 21.9, 21.6, 21.2, 21.3, 21.5, 20.3  
 1955–2000 (5-yr intervals): 20.2, 20.2, 20.6, 20.8, 21.1, 22.0, 23.3, 23.9, 24.5, 25.1

- How much faith do you place in this prediction? Explain.
- Do you think your model would produce an accurate prediction about your grandchildren, say, 50 years from now? Explain.

**30. Unwed births.** The National Center for Health Statistics reported the data below, showing the percentage of all births that are to unmarried women for selected years

between 1980 and 1998. Create a model that describes this trend. Justify decisions you make about how to best use these data.

Year	1980	1985	1990	1991	1992	1993	1994	1995	1996	1997	1998
%	18.4	22.0	28.0	29.5	30.1	31.0	32.6	32.2	32.4	32.4	32.8

**T 31. Life Expectancy 2004.** Data from the World Bank for 26 Western Hemisphere countries can be used to examine the association between female *Life Expectancy* and the average *Number of Children* women give birth to (<http://devdata.worldbank.org/data-query/>).

Country	Births/Woman	Life Exp.	Country	Births/Woman	Life Exp.
Argentina	2.3	74.6	Guatemala	4.4	67.6
Bahamas	2.3	70.5	Honduras	3.6	68.2
Barbados	1.7	75.4	Jamaica	2.4	70.8
Belize	3.0	71.9	Mexico	2.2	75.1
Bolivia	3.7	64.5	Nicaragua	3.2	70.1
Brazil	2.3	70.9	Panama	2.6	75.1
Canada	1.5	79.8	Paraguay	3.7	71.2
Chile	2.0	78.0	Peru	2.8	70.4
Colombia	2.4	72.6	Puerto Rico	1.9	77.5
Costa Rica	24.9	78.7	United States	2.0	77.4
Dominican Republic	2.8	67.8	Uruguay	2.1	75.2
Ecuador	2.7	74.5	Venezuela	2.7	73.7
El Salvador	2.8	71.1	Virgin Islands	2.2	78.6

- Create a scatterplot relating these two variables, and describe the association.
- Are there any countries that do not seem to fit the overall pattern?
- Find the correlation, and interpret the value of  $R^2$ .
- Find the equation of the regression line.
- Is the line an appropriate model? Describe what you see in the residuals plot.
- Interpret the slope and the  $y$ -intercept of the line.
- If government leaders wanted to increase life expectancy in their country, should they encourage women to have fewer children? Explain.

**T 32. Tour de France 2007.** We met the Tour de France data set in Chapter 2 (in Just Checking). One hundred years ago, the fastest rider finished the course at an average speed of about 25.3 kph (around 15.8 mph). In 2005, Lance Armstrong averaged 41.65 kph (25.88 mph) for the fastest average winning speed in history.

- Make a scatterplot of *Avg Speed* against *Year*. Describe the relationship of *Avg Speed* by *Year*, being careful to point out any unusual features in the plot.
- Find the regression equation of *Avg Speed* on *Year*.
- Are the conditions for regression met? Comment.

**T 33. Inflation 2006.** The Consumer Price Index (CPI) tracks the prices of consumer goods in the United States, as shown in the table on the next page (<ftp://ftp.bis.gov>). It



indicates, for example, that the average item costing \$17.70 in 1926 cost \$201.60 in the year 2006.

Year	CPI	Year	CPI
1914	10.0	1962	30.2
1918	15.1	1966	32.4
1922	16.8	1970	38.8
1926	17.7	1974	49.3
1930	16.7	1978	65.2
1934	13.4	1982	96.5
1938	14.1	1986	109.6
1942	16.3	1990	130.7
1946	19.5	1994	148.2
1950	24.1	1998	163.0
1954	26.9	2002	179.9
1958	28.9	2006	201.6

- Make a scatterplot showing the trend in consumer prices. Describe what you see.
- Be an economic forecaster: Project increases in the cost of living over the next decade. Justify decisions you make in creating your model.

- T 34. Second stage 2007.** Look once more at the data from the Tour de France. In Exercise 32 we looked at the whole history of the race, but now let's consider just the post-World War II era.
- Find the regression of *Avg Speed* by *Year* only for years from 1947 to the present. Are the conditions for regression met?
  - Interpret the slope.
  - In 1979 Bernard Hinault averaged 39.8 kph, while in 2005 Lance Armstrong averaged 41.65 kph. Which was the more remarkable performance and why?



### JUST CHECKING Answers

- Not high leverage, not influential, large residual
- High leverage, not influential, small residual
- High leverage, influential, not large residual

# Re-expressing Data: Get It Straight!



**A S** **Activity: Re-expressing Data.** Should you re-express data? Actually, you already do.

**H**ow fast can you go on a bicycle? If you measure your speed, you probably do it in miles per hour or kilometers per hour. In a 12-mile-long time trial in the 2005 Tour de France, Dave Zabriskie *averaged* nearly 35 mph (54.7 kph), beating Lance Armstrong by 2 seconds. You probably realize that's a tough act to follow. It's fast. You can tell that at a glance because you have no trouble thinking in terms of distance covered per time.

OK, then, if you averaged 12.5 mph (20.1 kph) for a mile *run*, would *that* be fast? Would it be fast for a 100-m dash? Even if you run the mile often, you probably have to stop and calculate. Running a mile in under 5 minutes (12 mph) is fast. A mile at 16 mph would be a world record (that's a 3-minute, 45-second mile). There's no single *natural* way to measure speed. Sometimes we use time over distance; other times we use the *reciprocal*, distance over time. Neither one is *correct*. We're just used to thinking that way in each case.

So, how does this insight help us understand data? All quantitative data come to us measured in some way, with units specified. But maybe those units aren't the best choice. It's not that meters are better (or worse) than fathoms or leagues. What we're talking about is re-expressing the data another way by applying a function, such as a square root, log, or reciprocal. You already use some of them, even though you may not know it. For example, the Richter scale of earthquake strength (logs), the decibel scale for sound intensity (logs), the *f*/stop scale for camera aperture openings (squares), and the gauges of shotguns (square roots) all include simple functions of this sort.

Why bother? As with speeds, some expressions of the data may be easier to think about. And some may be much easier to analyze with statistical methods. We've seen that symmetric distributions are easier to summarize and straight scatterplots are easier to model with regressions. We often look to re-express our data if doing so makes them more suitable for our methods.

Scan through any Physics book. Most equations have powers, reciprocals, or logs.

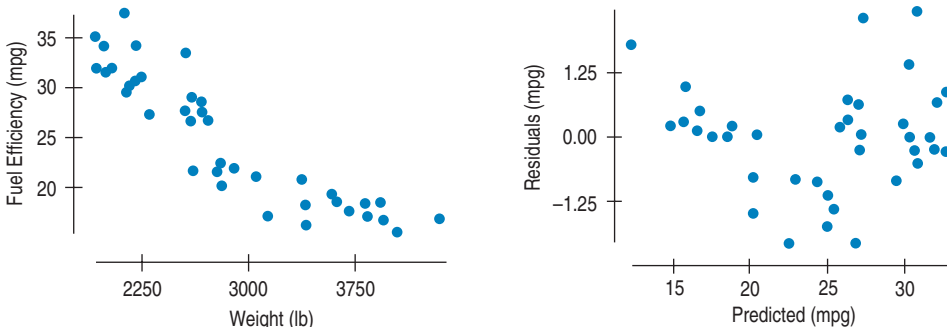
## Straight to the Point

We know from common sense and from physics that heavier cars need more fuel, but exactly how does a car's weight affect its fuel efficiency? Here are the

scatterplot of *Weight* (in pounds) and *Fuel Efficiency* (in miles per gallon) for 38 cars, and the residuals plot:

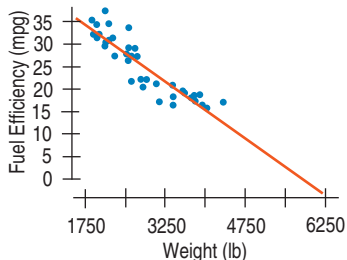
**FIGURE 10.1**

**Fuel Efficiency (mpg) vs. Weight for 38 cars as reported by Consumer Reports.** The scatterplot shows a negative direction, roughly linear shape, and strong relationship. However, the residuals from a regression of Fuel Efficiency on Weight reveal a bent shape when plotted against the predicted values. Looking back at the original scatterplot, you may be able to see the bend.



Hmm . . . Even though  $R^2$  is 81.6%, the residuals don't show the random scatter we were hoping for. The shape is clearly bent. Looking back at the first scatterplot, you can probably see the slight bending. Think about the regression line through the points. How heavy would a car have to be to have a predicted gas mileage of 0? It looks like the *Fuel Efficiency* would go negative at about 6000 pounds. A Hummer H2 weighs about 6400 pounds. The H2 is hardly known for fuel efficiency, but it does get more than the *minus 5 mpg* this regression predicts. Extrapolation is always dangerous, but it's more dangerous the more the model is wrong, because wrong models tend to do even worse the farther you get from the middle of the data.

The bend in the relationship between *Fuel Efficiency* and *Weight* is the kind of failure to satisfy the conditions for an analysis that we can repair by re-expressing the data. Instead of looking at miles per gallon, we could take the reciprocal and work with gallons per hundred miles.<sup>1</sup>



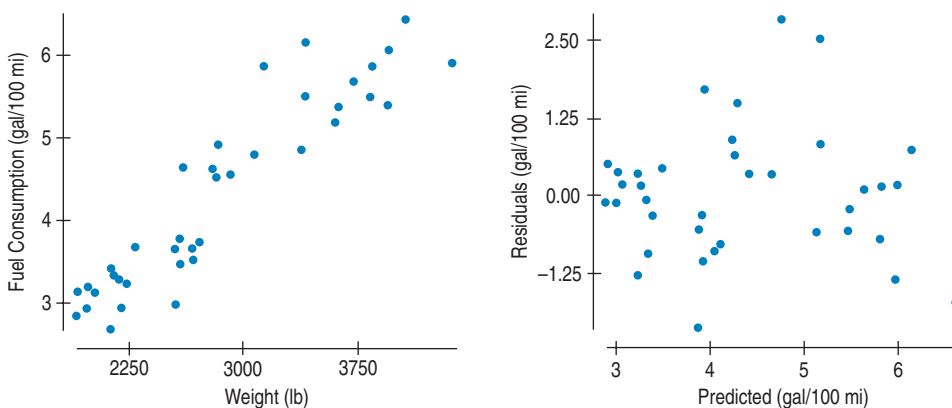
**FIGURE 10.2**

Extrapolating the regression line gives an absurd answer for vehicles that weigh as little as 6000 pounds.

**“Gallons per hundred miles—what an absurd way to measure fuel efficiency! Who would ever do it that way?”** Not all re-expressions are easy to understand, but in this case the answer is “Everyone except U.S. drivers.” Most of the world measures fuel efficiency in liters per 100 kilometers (L/100 km). This is the same reciprocal form (fuel amount per distance driven) and differs from gallons per 100 miles only by a constant multiple of about 2.38. It has been suggested that most of the world says, “I’ve got to go 100 km; how much gas do I need?” But Americans say, “I’ve got 10 gallons in the tank. How far can I drive?” In much the same way, re-expressions “think” about the data differently but don’t change what they mean.

**FIGURE 10.3**

The reciprocal ( $1/y$ ) is measured in gallons per mile. Gallons per 100 miles gives more meaningful numbers. The reciprocal is more nearly linear against Weight than the original variable, but the re-expression changes the direction of the relationship. The residuals from the regression of Fuel Consumption (gal/100 mi) on Weight show less of a pattern than before.



<sup>1</sup> Multiplying by 100 to get gallons per 100 miles simply makes the numbers easier to think about: You might have a good idea of how many gallons your car needs to drive 100 miles, but probably a much poorer sense of how much gas you need to go just 1 mile.

The direction of the association is positive now, since we’re measuring gas consumption and heavier cars consume more gas per mile. The relationship is much straighter, as we can see from a scatterplot of the regression residuals.

This is more the kind of boring residuals plot (no direction, no particular shape, no outliers, no bends) that we hope to see, so we have reason to think that the Straight Enough Condition is now satisfied. Now here’s the payoff: What does the reciprocal model say about the Hummer? The regression line fit to *Fuel Consumption vs. Weight* predicts somewhere near 9.7 for a car weighing 6400 pounds. What does this mean? It means the car is predicted to use 9.7 gallons for every 100 miles, or in other words,

$$\frac{100 \text{ miles}}{9.7 \text{ gallons}} = 10.3 \text{ mpg.}$$

That’s a much more reasonable prediction and very close to the reported value of 11.0 miles per gallon (of course, *your* mileage may vary . . .).

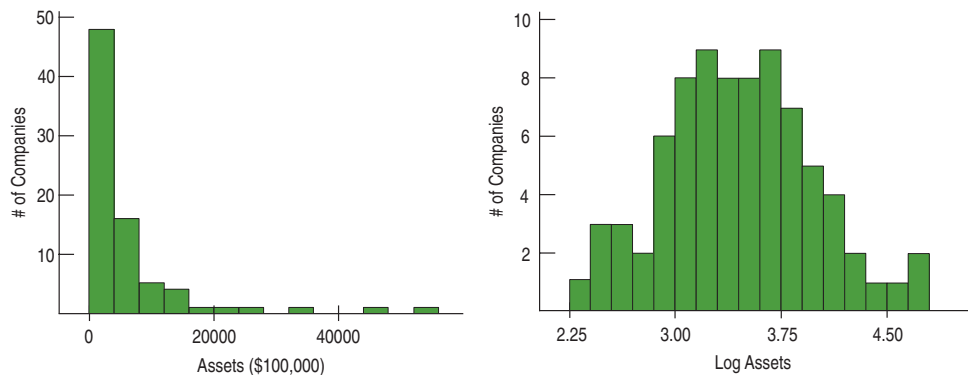
## Goals of Re-expression

We re-express data for several reasons. Each of these goals helps make the data more suitable for analysis by our methods.

### GOAL 1

*Make the distribution of a variable (as seen in its histogram, for example) more symmetric.* It’s easier to summarize the center of a symmetric distribution, and for nearly symmetric distributions, we can use the mean and standard deviation. If the distribution is unimodal, then the resulting distribution may be closer to the Normal model, allowing us to use the 68–95–99.7 Rule.

Here are a histogram, quite skewed, showing the *Assets* of 77 companies selected from the Forbes 500 list (in \$100,000) and the more symmetric histogram after taking logs.



**FIGURE 10.4**

*The distribution of the Assets of large companies is skewed to the right. Data on wealth often look like this. Taking logs makes the distribution more nearly symmetric.*

### GOAL 2

*Make the spread of several groups (as seen in side-by-side boxplots) more alike, even if their centers differ.* Groups that share a common spread are easier to compare. We’ll see methods later in the book that can be applied only to groups with

**WHO** 77 large companies  
**WHAT** Assets, sales, and market sector  
**UNITS** \$100,000  
**HOW** Public records  
**WHEN** 1986  
**WHY** By *Forbes* magazine in reporting on the Forbes 500 for that year

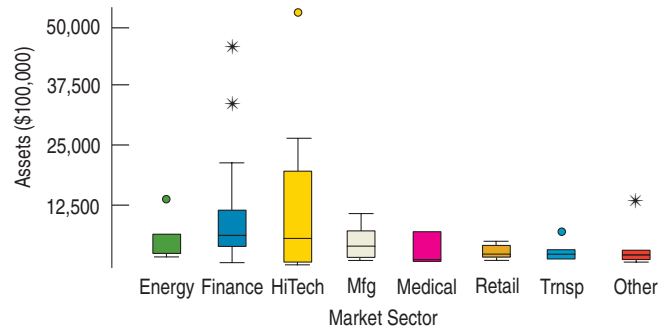
**A S** **Simulation: Re-expression in Action.** Slide the re-expression power and watch the histogram change.

a common standard deviation. We saw an example of re-expression for comparing groups with boxplots in Chapter 5.

Here are the *Assets* of these companies by *Market Sector*:

**FIGURE 10.5**

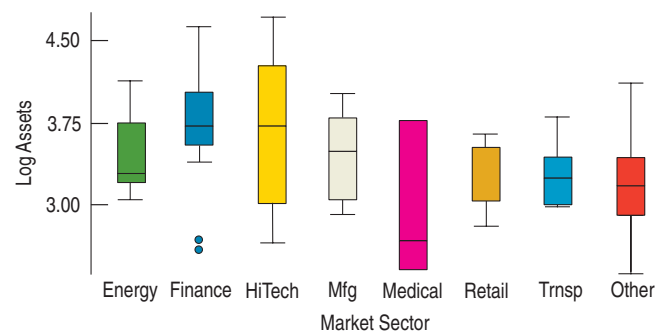
**Assets of large companies by Market Sector.** It's hard to compare centers or spreads, and there seem to be a number of high outliers.



Taking logs makes the individual boxplots more symmetric and gives them spreads that are more nearly equal.

**FIGURE 10.6**

After re-expressing by logs, it's much easier to compare across market sectors. The boxplots are more nearly symmetric, most have similar spreads, and the companies that seemed to be outliers before are no longer extraordinary. Two new outliers have appeared in the finance sector. They are the only companies in that sector that are not banks. Perhaps they don't belong there.



Doing this makes it easier to compare assets across market sectors. It can also reveal problems in the data. Some companies that looked like outliers on the high end turned out to be more typical. But two companies in the finance sector now stick out. Unlike the rest of the companies in that sector, they are not banks. They may have been placed in the wrong sector, but we couldn't see that in the original data.

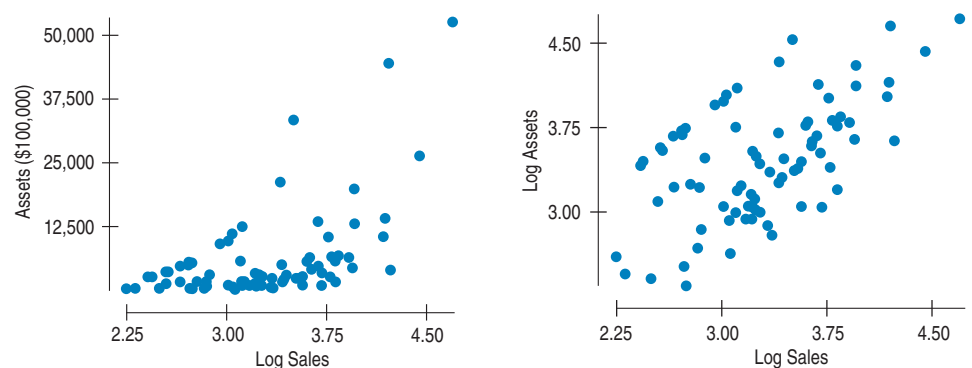
### GOAL 3

**Make the form of a scatterplot more nearly linear.** Linear scatterplots are easier to model. We saw an example of scatterplot straightening in Chapter 7. The greater value of re-expression to straighten a relationship is that we can fit a linear model once the relationship is straight.

Here are *Assets* of the companies plotted against the logarithm of *Sales*, clearly bent. Taking logs makes things much more linear.

**FIGURE 10.7**

*Assets* vs. *log Sales* shows a positive association (bigger sales go with bigger assets) but a bent shape. Note also that the points go from tightly bunched at the left to widely scattered at the right; the plot "thickens." In the second plot, *log Assets* vs. *log Sales* shows a clean, positive, linear association. And the variability at each value of *x* is about the same.



### GOAL 4

*Make the scatter in a scatterplot spread out evenly rather than thickening at one end.* Having an even scatter is a condition of many methods of Statistics, as we'll see in later chapters. This goal is closely related to Goal 2, but it often comes along with Goal 3. Indeed, a glance back at the scatterplot (Figure 10.7) shows that the plot for *Assets* is much more spread out on the right than on the left, while the plot for  $\log Assets$  has roughly the same variation in  $\log Assets$  for any  $x$ -value.

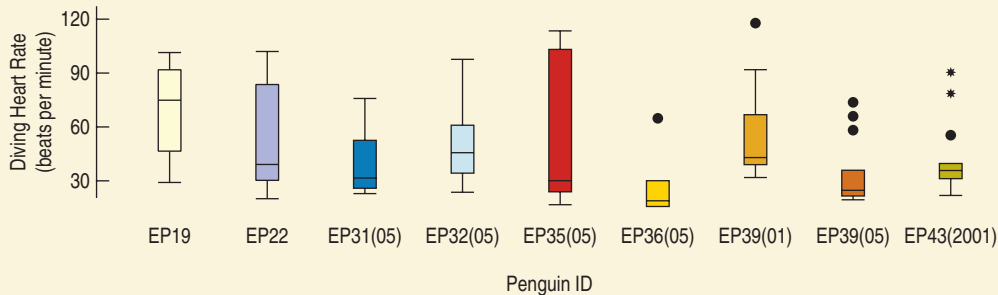
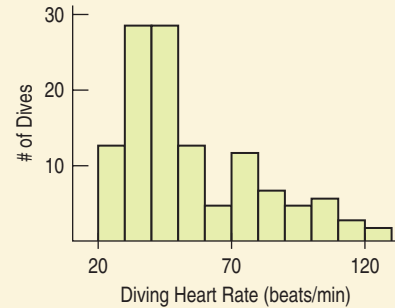
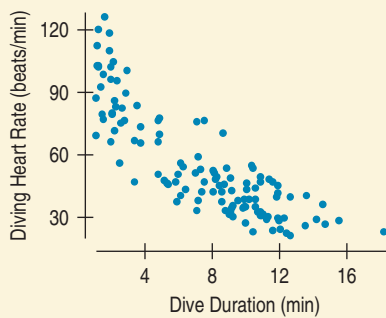
#### FOR EXAMPLE

#### Recognizing when a re-expression can help

In Chapter 9, we saw the awesome ability of emperor penguins to slow their heart rates while diving. Here are three displays relating to the diving heart rates:

(The boxplots show the diving heart rates for each of the 9 penguins whose dives were tracked. The names are those given by the researchers; EP = emperor penguin.)

**Question:** What features of each of these displays suggest that a re-expression might be helpful?



The scatterplot shows a curved relationship, concave upward, between the duration of the dives and penguins' heart rates. Re-expressing either variable may help to straighten the pattern.

The histogram of heart rates is skewed to the high end. Re-expression often helps to make skewed distributions more nearly symmetric.

The boxplots each show skewness to the high end as well. The medians are low in the boxes, and several show high outliers.

## The Ladder of Powers

**AS** **Activity: Re-expression in Action** Here's the animated version of the Ladder of Powers. Slide the power and watch the change.

How can we pick a re-expression to use? Some kinds of data favor certain re-expressions. But even starting from a suggested one, it's always a good idea to look around a bit. Fortunately, the re-expressions line up in order, so it's easy to slide up and down to find the best one. The trick is to choose our re-expressions from a simple family that includes the most common ways to re-express data. More important, the members of the family line up in order, so that the farther you move away from the original data (the "1" position), the greater is the effect on the data. This fact lets you search systematically for a re-expression that

TI-*n*spire

**Re-expression.** See a curved relationship become straighter with each step on the Ladder of Powers.

works, stepping a bit farther from “1” or taking a step back toward “1” as you see the results.

Where to start? It turns out that certain kinds of data are more likely to be helped by particular re-expressions. Knowing that gives you a good place to start your search for a re-expression. We call this collection of re-expressions the **Ladder of Powers**.

Power	Name	Comment
2	The square of the data values, $y^2$ .	Try this for unimodal distributions that are skewed to the left.
1	The raw data—no change at all. This is “home base.” The farther you step from here up or down the ladder, the greater the effect.	Data that can take on both positive and negative values with no bounds are less likely to benefit from re-expression.
1/2	The square root of the data values, $\sqrt{y}$ .	Counts often benefit from a square root re-expression. For counted data, start here.
“0”	Although mathematicians define the “0-th” power differently, <sup>2</sup> for us the place is held by the logarithm. You may feel uneasy about logarithms. Don’t worry; the computer or calculator does the work. <sup>3</sup>	Measurements that cannot be negative, and especially values that grow by percentage increases such as salaries or populations, often benefit from a log re-expression. When in doubt, start here. If your data have zeros, try adding a small constant to all values before finding the logs.
-1/2	The (negative) reciprocal square root, $-1/\sqrt{y}$ .	An uncommon re-expression, but sometimes useful. Changing the sign to take the <i>negative</i> of the reciprocal square root preserves the direction of relationships, making things a bit simpler.
-1	The (negative) reciprocal, $-1/y$ .	Ratios of two quantities (miles per hour, for example) often benefit from a reciprocal. (You have about a 50–50 chance that the original ratio was taken in the “wrong” order for simple statistical analysis and would benefit from re-expression.) Often, the reciprocal will have simple units (hours per mile). Change the sign if you want to preserve the direction of relationships. If your data have zeros, try adding a small constant to all values before finding the reciprocal.



## JUST CHECKING

1. You want to model the relationship between the number of birds counted at a nesting site and the temperature (in degrees Celsius). The scatterplot of counts vs. temperature shows an upwardly curving pattern, with more birds spotted at higher temperatures. What transformation (if any) of the bird counts might you start with?
2. You want to model the relationship between prices for various items in Paris and in Hong Kong. The scatterplot of Hong Kong prices vs. Parisian prices shows a generally straight pattern with a small amount of scatter. What transformation (if any) of the Hong Kong prices might you start with?
3. You want to model the population growth of the United States over the past 200 years. The scatterplot shows a strongly upwardly curved pattern. What transformation (if any) of the population might you start with?

<sup>2</sup> You may remember that for any nonzero number  $y$ ,  $y^0 = 1$ . This is not a very exciting transformation for data; every data value would be the same. We use the logarithm in its place.

<sup>3</sup> Your calculator or software package probably gives you a choice between “base 10” logarithms and “natural (base  $e$ )” logarithms. Don’t worry about that. It doesn’t matter at all which you use; they have exactly the same effect on the data. If you want to choose, base 10 logarithms can be a bit easier to interpret.

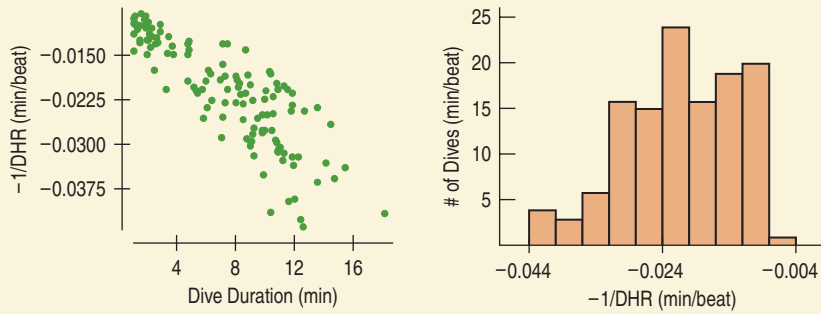
Scientific laws often include simple re-expressions. For example, in Psychology, Fechner’s Law states that sensation increases as the logarithm of stimulus intensity ( $S = k \log R$ ).

The Ladder of Powers orders the effects that the re-expressions have on data. If you try, say, taking the square roots of all the values in a variable and it helps, but not enough, then move farther down the ladder to the logarithm or reciprocal root. Those re-expressions will have a similar, but even stronger, effect on your data. If you go too far, you can always back up. But don’t forget—when you take a negative power, the *direction* of the relationship will change. That’s OK. You can always change the sign of the response variable if you want to keep the same direction. With modern technology, finding a suitable re-expression is no harder than the push of a button.

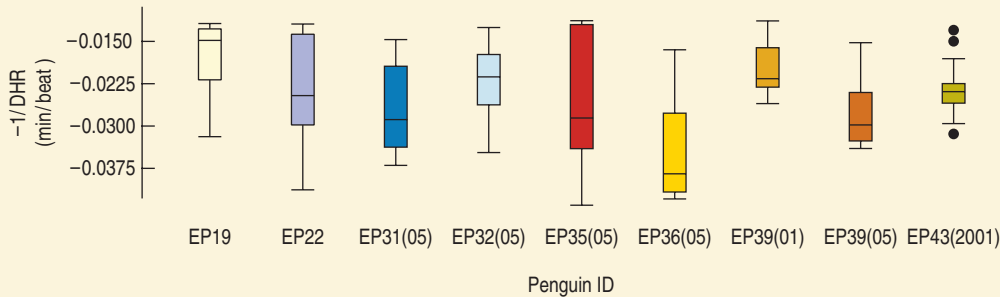
**FOR EXAMPLE**

**Trying a re-expression**

**Recap:** We’ve seen curvature in the relationship between emperor penguins’ diving heart rates and the duration of the dive. Let’s start the process of finding a good re-expression. Heart rate is in beats per minute; maybe heart “speed” in minutes per beat would be a better choice. Here are the corresponding displays for this reciprocal re-expression (as we often do, we’ve changed the sign to preserve the order of the data values):



**Question:** Were the re-expressions successful?



The scatterplot bends less than before, but now may be slightly concave downward. The histogram is now slightly skewed to the low end. Most of the boxplots have no outliers. These boxplots seem better than the ones for the raw heart rates. Overall, it looks like I may have moved a bit “too far” on the ladder of powers. Halfway between “1” (the original data) and “-1” (the reciprocal) is “0”, which represents the logarithm. I’d try that for comparison.

**STEP-BY-STEP EXAMPLE**

**Re-expressing to Straighten a Scatterplot**

Standard (monofilament) fishing line comes in a range of strengths, usually expressed as “test pounds.” Five-pound test line, for example, can be expected to withstand a pull of up to five pounds without breaking. The convention in selling fishing line is that the price of a spool doesn’t vary with strength. Instead, the length of line on the spool varies. Higher test pound line is thicker, though, so spools of fishing line hold about the same amount of material. Some spools hold line that is thinner and longer, some fatter and shorter. Let’s look at the *Length* and *Strength* of spools of monofilament line manufactured by the same company and sold for the same price at one store.



**Questions:** How are the *Length* on the spool and the *Strength* related? And what re-expression will straighten the relationship?

**THINK**

**Plan** State the problem.

**Variables** Identify the variables and report the W's.

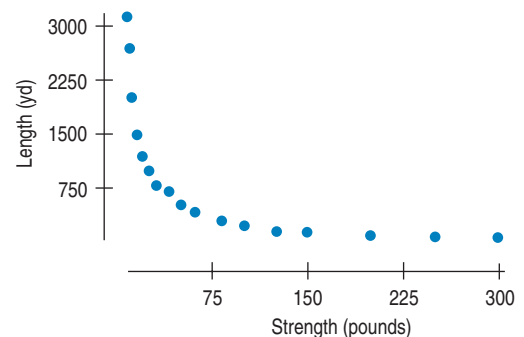
**Plot** Check that even if there is a curve, the overall pattern does not reach a minimum or maximum and then turn around and go back. An up-and-down curve can't be fixed by re-expression.

I want to fit a linear model for the length and strength of monofilament fishing line.

I have the *length* and "pound test" *strength* of monofilament fishing line sold by a single vendor at a particular store. Each case is a different strength of line, but all spools of line sell for the same price.

Let *Length* = length (in yards) of fishing line on the spool

*Strength* = the test strength (in pounds).



The plot shows a negative direction and an association that has little scatter but is not straight.

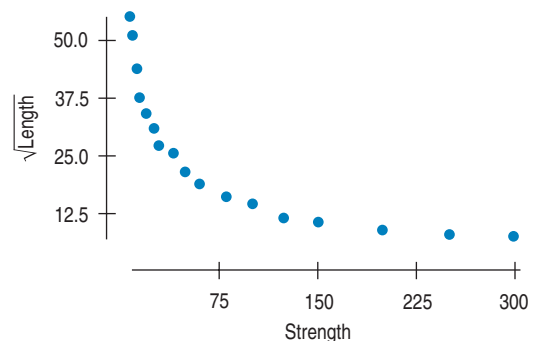
**SHOW**

**Mechanics** Try a re-expression.

The lesson of the Ladder of Powers is that if we're moving in the right direction but have not had sufficient effect, we should go farther along the ladder. This example shows improvement, but is still not straight.

(Because *Length* is an amount of something and cannot be negative, we probably should have started with logs. This plot is here in part to illustrate how the Ladder of Powers works.)

Here's a plot of the square root of *Length* against *Strength*:



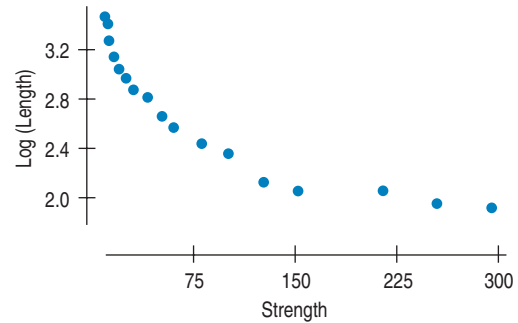
The plot is less bent, but still not straight.

Stepping from the  $1/2$  power to the “0” power, we try the logarithm of *Length* against *Strength*.

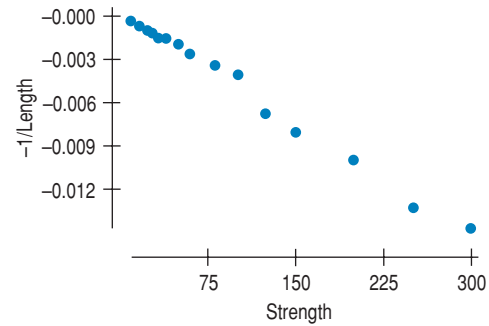
The straightness is improving, so we know we’re moving in the right direction. But since the plot of the logarithms is not yet straight, we know we haven’t gone far enough. To keep the direction consistent, change the sign and re-express to  $-1/Length$ .

We may have to choose between two adjacent re-expressions. For most data analyses, it really doesn’t matter which we choose.

The scatterplot of the logarithm of *Length* against *Strength* is even less bent:

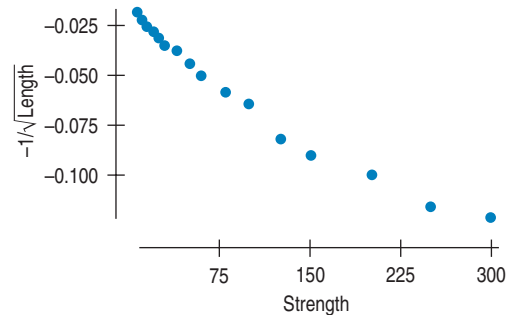


This is much better, but still not straight, so I’ll take another step to the “-1” power, or reciprocal.



Maybe now I moved too far along the ladder.

A half-step back is the  $-1/2$  power: the reciprocal square root.



**Conclusion** Specify your choice of re-expression. If there’s some natural interpretation (as for gallons per 100 miles), give that.

It’s hard to choose between the last two alternatives. Either of the last two choices is good enough. I’ll choose the  $-1/2$  power.

Now that the re-expressed data satisfy the Straight Enough Condition, we can fit a linear model by least squares. We find that

$$\frac{-1}{\sqrt{\widehat{Length}}} = -0.023 - 0.000373 \text{ Strength}.$$

We can use this model to predict the length of a spool of, say, 35-pound test line:

$$\frac{-1}{\sqrt{\widehat{Length}}} = -0.023 - 0.000373 \times 35 = -0.036$$

We could leave the result in these units ( $-1/\sqrt{\text{yards}}$ ). Sometimes the new units may be as meaningful as the original, but here we want to transform the predicted value back into yards. Fortunately, each of the re-expressions in the Ladder of Powers can be reversed.

To reverse the process, we first take the reciprocal:  $\sqrt{\widehat{Length}} = -1/(-0.036) = 27.778$ . Then squaring gets us back to the original units:

$$\widehat{Length} = 27.778^2 = 771.6 \text{ yards}.$$

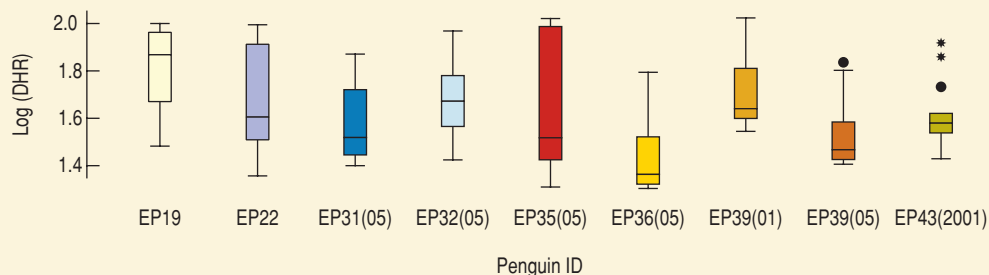
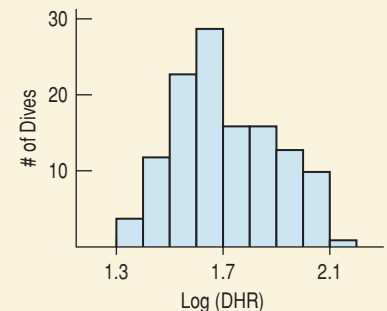
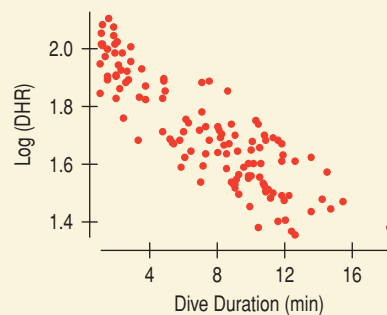
This may be the most painful part of the re-expression. Getting back to the original units can sometimes be a little work. Nevertheless, it's worth the effort to always consider re-expression. Re-expressions extend the reach of all of your Statistics tools by helping more data to satisfy the conditions they require. Just think how much more useful this course just became!

## FOR EXAMPLE

### Comparing re-expressions

**Recap:** We've concluded that in trying to straighten the relationship between *Diving Heart Rate* and *Dive Duration* for emperor penguins, using the reciprocal re-expression goes a bit "too far" on the ladder of powers. Now we try the logarithm. Here are the resulting displays:

**Questions:** Comment on these displays. Now that we've looked at the original data (rung 1 on the Ladder), the reciprocal (rung -1), and the logarithm (rung 0), which re-expression of *Diving Heart Rate* would you choose?



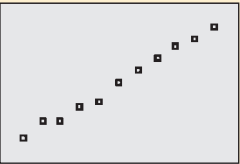
The scatterplot is now more linear and the histogram is symmetric. The boxplots are still a bit skewed to the high end, but less so than for the original *Diving Heart Rate* values. We don't expect real data to cooperate perfectly, and the logarithm seems like the best compromise re-expression, improving several different aspects of the data.

## TI Tips

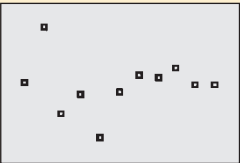
## Re-expressing data to achieve linearity



```
log(LTUIT)→L1
(3.815976001 3....
```



```
LinReg
y=a+bx
a=3.815541881
b=.0175535352
r²=.9908736906
r=.9954263863
```



```
Y1(11)
4.008630769
10^(Ans)
10200.71864
```

Let's revisit the Arizona State tuition data. Recall that back in Chapter 8 when we tried to fit a linear model to the yearly tuition costs, the residuals plot showed a distinct curve. Residuals are high (positive) at the left, low in the middle of the decade, and high again at the right.

This curved pattern indicates that data re-expression may be in order. If you have no clue what re-expression to try, the Ladder of Powers may help. We just used that approach in the fishing line example. Here, though, we can play a hunch. It is reasonable to suspect that tuition increases at a relatively consistent percentage year by year. This suggests that using the logarithm of tuition may help.

- Tell the calculator to find the logs of the tuitions, and store them as a new list. Remember that you must import the name `TUIT` from the `LIST NAMES` menu. The command is `log(LTUIT) STO L1`.
- Check the scatterplot for the re-expressed data by changing your `STATPLOT` specifications to `Xlist:YR` and `Ylist:L1`. (Don't forget to use `9: ZoomStat` to resize the window properly.)

The new scatterplot looks quite linear, but it's really the residuals plot that will tell the story. Remember that the TI automatically finds and stores the residuals whenever you ask it to calculate a regression.

- Perform the regression for the logarithm of *tuition* vs. *year* with the command `LinReg(a+bx) L1, Y1`. That both creates the residuals and reports details about the model (storing the equation for later use).
- Now that the residuals are stored in `RESID`, set up a new scatterplot, this time specifying `Xlist:YR` and `Ylist:RESID`.

While the residuals for the second and fifth years are comparatively large, the curvature we saw above is gone. The pattern in these residuals seem essentially horizontal and random. This re-expressed model is probably more useful than the original linear model.

Do you know what the model's equation is? Remember, it involves a log re-expression. The calculator does not indicate that; be sure to *Think* when you write your model!

$$\log \widehat{tuit} = 3.816 + 0.018 yr$$

And you have to *Think* some more when you make an estimate using the calculator's equation. Notice that this model does not actually predict tuition; rather, it predicts the *logarithm* of the tuition.

For example, to estimate the 2001 tuition we must first remember that in entering our data we designated 1990 as year 0. That means we'll use 11 for the year 2001 and evaluate `Y1(11)`.

No, we're not predicting the tuition to be \$4! That's the log of the estimated tuition. Since logarithms are exponents,  $\log(\widehat{tuit}) = 4$  means  $\widehat{tuit} = 10^4$ , or about \$10,000. When you are working with models that involve re-expressions, you'll often need to "backsolve" like this to find the correct predictions.

## Plan B: Attack of the Logarithms

The Ladder of Powers is often successful at finding an effective re-expression. Sometimes, though, the curvature is more stubborn, and we're not satisfied with the residual plots. What then?

When none of the data values is zero or negative, logarithms can be a helpful ally in the search for a useful model. Try taking the logs of both the  $x$ - and  $y$ -variables. Then re-express the data using some combination of  $x$  or  $\log(x)$  vs.  $y$  or  $\log(y)$ . You may find that one of these works pretty well.

Model Name	$x$ -axis	$y$ -axis	Comment
Exponential	$x$	$\log(y)$	This model is the "0" power in the ladder approach, useful for values that grow by percentage increases.
Logarithmic	$\log(x)$	$y$	A wide range of $x$ -values, or a scatterplot descending rapidly at the left but leveling off toward the right, may benefit from trying this model.
Power	$\log(x)$	$\log(y)$	The Goldilocks model: When one of the ladder's powers is too big and the next is too small, this one may be just right.

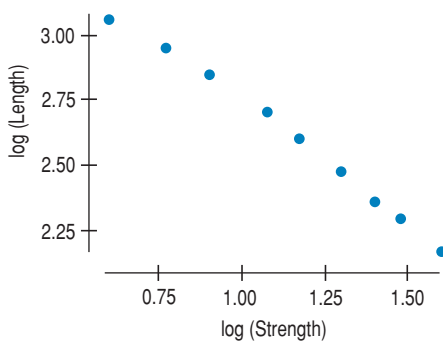


FIGURE 10.8

Plotting  $\log(\text{Length})$  against  $\log(\text{Strength})$  gives a straighter shape.

When we tried to model the relationship between the length of fishing line and its strength, we were torn between the “ $-1$ ” power and the “ $-1/2$ ” power. The first showed slight upward curvature, and the second downward. Maybe there's a better power between those values.

The scatterplot shows what happens when we graph the logarithm of *Length* against the logarithm of *Strength*. Technology reveals that the equation of our log-log model is

$$\widehat{\log(\text{Length})} = 4.49 - 1.08 \log(\text{Strength}).$$

It's interesting that the slope of this line ( $-1.08$ ) is a power<sup>4</sup> we didn't try. After all, the ladder can't have every imaginable rung.

A warning, though! Don't expect to be able to straighten every curved scatterplot you find. It may be that there just isn't a very effective re-expression to be had. You'll certainly encounter situations when nothing seems to work the way you wish it would. Don't set your sights too high—you won't find a perfect model. Keep in mind: We seek a *useful* model, not perfection (or even “the best”).

### TI Tips



### Using logarithmic re-expressions

In Chapter 7 we looked at data showing the relationship between the  $f$ /stop of a camera's lens and its shutter speed. Let's use the attack of the logarithms to model this situation.

<b>Shutter speed:</b>	1/1000	1/500	1/250	1/125	1/60	1/30	1/15	1/8
<b>f/stop:</b>	2.8	4	5.6	8	11	16	22	32

- Enter these data into your calculator, shutter *speed* in L1 and *f/stop* in L2.
- Create the scatterplot with  $\text{Xlist:L1}$  and  $\text{Ylist:L2}$ . See the curve?

<sup>4</sup> For logarithms,  $-1.08 \log(\text{Strength}) = \log(\text{Strength}^{-1.08})$ .

```
log(L1)→L3
(-3 -2.69897000...
log(L2)→L4
(.4471580313 .6...
```



```
LinReg
y=a+bx
a=1.93880413
b=.4969548956
r²=.9993420212
r=.9996709565
```

- Find the logarithms of each variable's values. Keep track of where you store everything so you don't get confused! We put  $\log(\text{speed})$  in **L3** and  $\log(f/\text{stop})$  in **L4**.
- Make three scatterplots:
  - $f/\text{stop}$  vs.  $\log(\text{speed})$  using **Xlist:L3** and **Ylist:L2**
  - $\log(f/\text{stop})$  vs.  $\text{speed}$  using **Xlist:L1** and **Ylist:L4**
  - $\log(f/\text{stop})$  vs.  $\log(\text{speed})$  using **Xlist:L3** and **Ylist:L4**
- Pick your favorite. We liked  $\log(f/\text{stop})$  vs.  $\log(\text{speed})$  a lot! It appears to be very straight. (Don't be misled—this is a situation governed by the laws of Physics. Real data are not so cooperative. Don't expect to achieve this level of perfection often!)
- Remember that before you check the residuals plot, you first have to calculate the regression. In this situation all the errors in the residuals are just round-off errors in the original  $f/\text{stops}$ .
- Use your regression to write the equation of the model. Remember: The calculator does not know there were logarithms involved. You have to Think about that to be sure you write your model correctly.<sup>5</sup>

$$\log(\widehat{f/\text{stop}}) = 1.94 + 0.497\log(\text{speed})$$

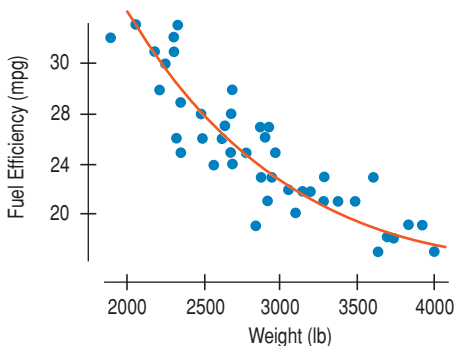
## Why Not Just Use a Curve?

When a clearly curved pattern shows up in the scatterplot, why not just fit a curve to the data? We saw earlier that the association between the *Weight* of a car and its *Fuel Efficiency* was not a straight line. Instead of trying to find a way to straighten the plot, why not find a curve that seems to describe the pattern well?

We can find “curves of best fit” using essentially the same approach that led us to linear models. You won't be surprised, though, to learn that the mathematics and the calculations are considerably more difficult for curved models. Many calculators and computer packages do have the ability to fit curves to data, but this approach has many drawbacks.

Straight lines are easy to understand. We know how to think about the slope and the  $y$ -intercept, for example. We often want some of the other benefits mentioned earlier, such as making the spread around the model more nearly the same everywhere. In later chapters you will learn more advanced statistical methods for analyzing linear associations.

We give all of that up when we fit a model that is not linear. For many reasons, then, it is usually better to re-express the data to straighten the plot.



### TI Tips

### Some shortcuts to avoid

Your calculator offers many regression options in the **STAT CALC** menu. There are three that automate fitting simple re-expressions of  $y$  or  $x$ :

- **9: LnReg**—fits a logarithmic model ( $\hat{y} = a + b\ln x$ )

<sup>5</sup> See the slope, 0.497? Just about 0.5. That's because the actual relationship involves the square root of shutter speeds. Technically the  $f/\text{stop}$  listed as 2.8 should be  $2\sqrt{2} \approx 2.8284$ . Rounding off to 2.8 makes sense for photographers, but it's what led to the minor errors you saw in the residuals plot.

- **0:ExpReg**—fits an exponential model ( $\hat{y} = ab^x$ )
- **A:PowReg**—fits a power model ( $\hat{y} = ax^b$ )

In addition, the calculator offers two other functions:

- **5:QuadReg**—fits a quadratic model ( $\hat{y} = ax^2 + bx + c$ )
- **6:CubicReg**—fits a cubic model ( $\hat{y} = ax^3 + bx^2 + cx + d$ )

These two models have a form we haven't seen, with several  $x$ -terms. Because  $x$ ,  $x^2$ , and  $x^3$  are likely to be highly correlated with each other, the quadratic and cubic models are almost sure to be unreliable to fit, difficult to understand, and dangerous to use for predictions even slightly outside the range of the data. We recommend that you be very wary of models of this type.

Let's try out one of the calculator shortcuts; we'll use the Arizona State tuition data. (For the last time, we promise!) This time, instead of re-expressing *tuition* to straighten the scatterplot, we'll have the calculator do more of the work.

Which model should you use? You could always just play hit-and-miss, but knowing something about the data can save a lot of time. If tuition increases by a consistent percentage each year, then the growth is exponential.

- Choose the exponential model, and specify your variables by importing **YR** and **TUIT** from the list names menu. And, because you'll want to graph the curve later, save its equation by adding **Y1** (from **VARS**, **Y-VARS**, **Function**) to create the command **ExpReg YR, LTUIT, Y1**.
- Set up the scatterplot. **ZoomStat** should show you the curve too.
- Graph the residuals plot.

This all looks very good.  $R^2$  is high, the curve appears to fit the points quite well, and the residuals plot is acceptably random.

The equation of the model is  $\widehat{tuit} = 6539.46(1.041^{year})$ .

Notice that this is the same residuals plot we saw when we re-expressed the data and fit a line to the logarithm of *tuition*. That's because what the calculator just did is mathematically the very same thing. This new equation may look different, but it is equivalent to our earlier model  $\log \widehat{tuit} = 3.816 + 0.018 \text{ year}$ .

Not easy to see that, is it? Here's how it works:

Initially we used a logarithmic re-expression to create a linear model:

$$\log \hat{y} = a + bx$$

Rewrite that equation in exponential form:

$$\hat{y} = 10^{a+bx}$$

Simplify, using the laws of exponents:

$$\hat{y} = 10^a(10^b)^x$$

Let  $10^a = a$  and  $10^b = b$  (different  $a$  and  $b$ !)

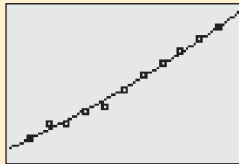
$$\hat{y} = ab^x$$

See? Your linear model created by logarithmic re-expression is the same as the calculator model created by **ExpReg**.

Three of the special TI functions correspond to a simple regression model involving re-expression. The calculator presents the results in an equation of a different form, but it doesn't actually fit that equation. Instead it is just doing the re-expression for you automatically.

```
ExpReg YR, LTUIT
Y1
```

```
ExpReg
Y=a*b^x
a=6539.459906
b=1.041246454
r^2=.9908736906
r=.9954263863
```



Here are the equivalent models for the two approaches.

Type of Model	Re-expression Equation	Calculator's Curve	
		Command	Equation
Logarithmic	$\hat{y} = a + b \log x$	LnReg	$\hat{y} = a + b \ln x$
Exponential	$\log \hat{y} = a + bx$	ExpReg	$\hat{y} = ab^x$
Power	$\log \hat{y} = a + b \log x$	PwrReg	$\hat{y} = ax^b$

Be careful. It may look like the calculator is fitting these equations to the data by minimizing the sum of squared residuals, but it isn't really doing that. It handles the residuals differently, and the difference matters. If you use a statistics program to fit an "exponential model," it will probably fit the exponential form of the equation and give you a different answer. So think of these TI functions as just shortcuts for fitting linear regressions to re-expressed versions of your data.

You've seen two ways to handle bent relationships:

- straighten the data, then fit a line, or
- use the calculator shortcut to create a curve.

Note that the calculator does not have a shortcut for every model you might want to use—models involving square roots or reciprocals, for instance. And remember: The calculator may be quick, but there are real advantages to finding *linear* models by actually re-expressing the data. That's the approach we strongly recommend you use.

## WHAT CAN GO WRONG?

### Occam's Razor

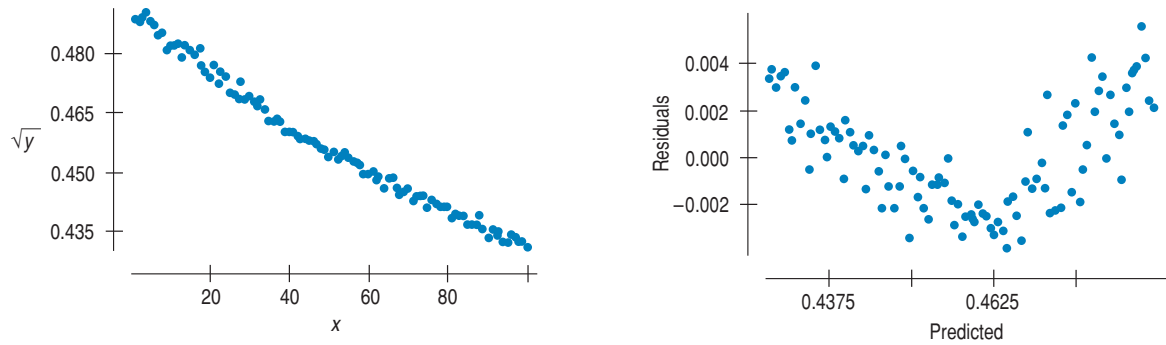
If you think that simpler explanations and simpler models are more likely to give a true picture of the way things work, then you should look for opportunities to re-express your data and simplify your analyses.

The general principle that simpler explanations are likely to be the better ones is known as Occam's Razor, after the English philosopher and theologian William of Occam (1284–1347).

- ▶ **Don't expect your model to be perfect.** In Chapter 6 we quoted statistician George Box: "All models are wrong, but some are useful." Be aware that the real world is a messy place and data can be uncooperative. Don't expect to find one elusive re-expression that magically irons out every kink in your scatterplot and produces perfect residuals. You aren't looking for the Right Model, because that mythical creature doesn't exist. Find a useful model and use it wisely.
- ▶ **Don't stray too far from the ladder.** It's wise not to stray too far from the powers that we suggest in the Ladder of Powers. Taking the  $y$ -values to an extremely high power may artificially inflate  $R^2$ , but it won't give a useful or meaningful model, so it doesn't really simplify anything. It's better to stick to powers between 2 and  $-2$ . Even in that range, you should prefer the simpler powers in the ladder to those in the cracks. A square root is easier to understand than the 0.413 power. That simplicity may compensate for a slightly less straight relationship.
- ▶ **Don't choose a model based on  $R^2$  alone.** You've tried re-expressing your data to straighten a curved relationship and found a model with a high  $R^2$ . Beware: That doesn't mean the pattern is straight now. On the next page is a plot of a relationship with an  $R^2$  of 98.3%.

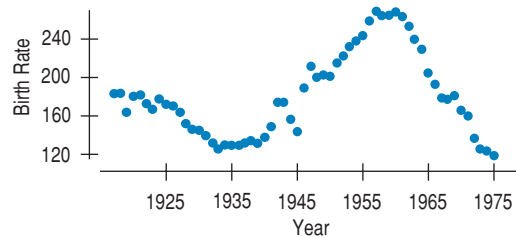
The  $R^2$  is about as high as we could ask for, but if you look closely, you'll see that there's a consistent bend. Plotting the residuals from the least squares line makes the bend much easier to see.





Remember the basic rule of data analysis: *Make a picture*. Before you fit a line, always look at the pattern in the scatterplot. After you fit the line, check for linearity again by plotting the residuals.

- ▶ **Beware of multiple modes.** Re-expression can often make a skewed unimodal histogram more nearly symmetric, but it cannot pull separate modes together. A suitable re-expression may, however, make the separation of the modes clearer, simplifying their interpretation and making it easier to separate them to analyze individually.
- ▶ **Watch out for scatterplots that turn around.** Re-expression can straighten many bent relationships but not those that go up and then down or down and then up. You should refuse to analyze such data with methods that require a linear form.



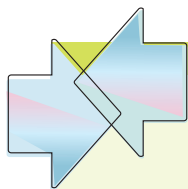
**FIGURE 10.9**

*The shape of the scatterplot of Birth Rates (births per 100,000 women) in the United States shows an oscillation that cannot be straightened by re-expressing the data.*

- ▶ **Watch out for negative data values.** It's impossible to re-express negative values by any power that is not a whole number on the Ladder of Powers or to re-express values that are zero for negative powers. Most statistics programs will just mark the result of trying to re-express such values "missing" if they can't be re-expressed. But that might mean that when you try a re-expression, you inadvertently lose a bunch of data values. The effect of that loss may be surprising and may substantially change your analysis. Because you are likely to be working with a computer package or calculator, take special care that you do not lose otherwise good data values when you choose a re-expression.

One possible cure for zeros and small negative values is to add a constant ( $\frac{1}{2}$  and  $\frac{1}{6}$  are often used) to bring all the data values above zero.

- ▶ **Watch for data far from 1.** Data values that are all very far from 1 may not be much affected by re-expression unless the range is very large. Re-expressing numbers between 1 and 100 will have a much greater effect than re-expressing numbers between 100,001 and 100,100. When all your data values are large (for example, working with years), consider subtracting a constant to bring them back near 1. (For example, consider "years since 1950" as an alternative variable for re-expression. Unless your data start at 1950, then avoid creating a zero by using "years since 1949.")



## CONNECTIONS

We have seen several ways to model or summarize data. Each requires that the data have a particular simple structure. We seek symmetry for summaries of center and spread and to use a Normal model. We seek equal variation across groups when we compare groups with boxplots or want to compare their centers. We seek linear shape in a scatterplot so that we can use correlation to summarize the scatter and regression to fit a linear model.

Data do often satisfy the requirements to use Statistics methods. But often they do not. Our choice is to stop with just displays, to use much more complex methods, or to re-express the data so that we can use the simpler methods we have developed.

In this fundamental sense, this chapter connects to everything we have done thus far and to all of the methods we will introduce throughout the rest of the book. Re-expression greatly extends the reach and applicability of all of these methods.



## WHAT HAVE WE LEARNED?

We've learned that when the conditions for regression are not met, a simple re-expression of the data may help. There are several reasons to consider a re-expression:

- ▶ To make the distribution of a variable more symmetric (as we saw in Chapter 5)
- ▶ To make the spread across different groups more similar
- ▶ To make the form of a scatterplot straighter
- ▶ To make the scatter around the line in a scatterplot more consistent

We've learned that when seeking a useful re-expression, taking logs is often a good, simple starting point. To search further, the Ladder of Powers or the log–log approach can help us find a good re-expression.

We've come to understand that our models won't be perfect, but that re-expression can lead us to a useful model.

### Terms

Re-expression

224. We re-express data by taking the logarithm, the square root, the reciprocal, or some other mathematical operation on all values of a variable.

Ladder of Powers

226. The Ladder of Powers places in order the effects that many re-expressions have on the data.

### Skills

THINK

- ▶ Recognize when a well-chosen re-expression may help you improve and simplify your analysis.
- ▶ Understand the value of re-expressing data to improve symmetry, to make the scatter around a line more constant, or to make a scatterplot more linear.
- ▶ Recognize when the pattern of the data indicates that no re-expression can improve the structure of the data.

SHOW

- ▶ Know how to re-express data with powers and how to find an effective re-expression for your data using your statistics software or calculator.

TELL

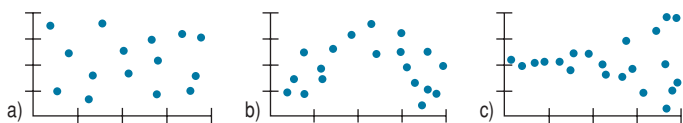
- ▶ Be able to reverse any of the common re-expressions to put a predicted value or residual back into the original units.
- ▶ Be able to describe a summary or display of a re-expressed variable, making clear how it was re-expressed and giving its re-expressed units.
- ▶ Be able to describe a regression model fit to re-expressed data in terms of the re-expressed variables.

## RE-EXPRESSION ON THE COMPUTER

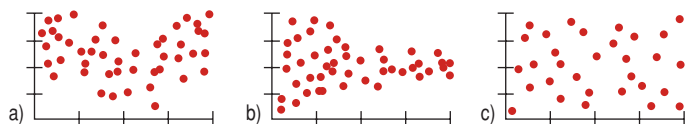
Computers and calculators make it easy to re-express data. Most statistics packages offer a way to re-express and compute with variables. Some packages permit you to specify the power of a re-expression with a slider or other moveable control, possibly while watching the consequences of the re-expression on a plot or analysis. This, of course, is a very effective way to find a good re-expression.

## EXERCISES

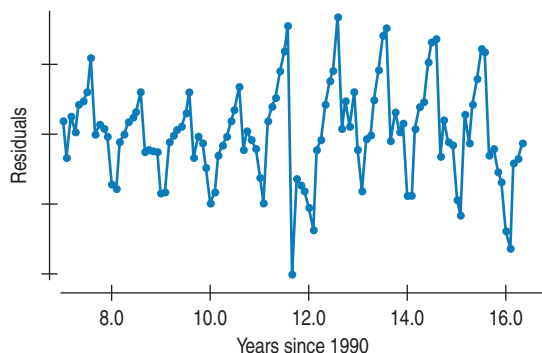
1. **Residuals.** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.



2. **Residuals.** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.

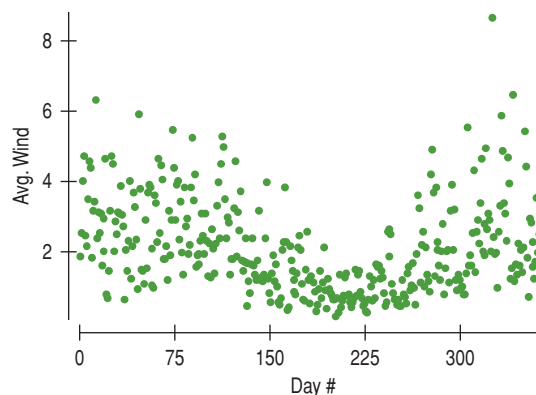


3. **Airline passengers revisited.** In Chapter 9, Exercise 9, we created a linear model describing the trend in the number of passengers departing from the Oakland (CA) airport each month since the start of 1997. Here's the residual plot, but with lines added to show the order of the values in time:



- a) Can you account for the pattern shown here?  
b) Would a re-expression help us deal with this pattern? Explain.

4. **Hopkins winds, revisited.** In Chapter 5, we examined the wind speeds in the Hopkins forest over the course of a year. Here's the scatterplot we saw then:



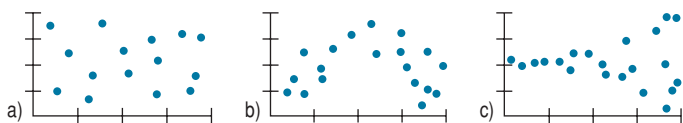
- a) Describe the pattern you see here.  
b) Should we try re-expressing either variable to make this plot straighter? Explain.
5. **Models.** For each of the models listed below, predict  $y$  when  $x = 2$ .
- |                                     |                                      |
|-------------------------------------|--------------------------------------|
| a) $\ln \hat{y} = 1.2 + 0.8x$       | d) $\hat{y} = 1.2 + 0.8 \ln x$       |
| b) $\sqrt{\hat{y}} = 1.2 + 0.8x$    | e) $\log \hat{y} = 1.2 + 0.8 \log x$ |
| c) $\frac{1}{\hat{y}} = 1.2 + 0.8x$ |                                      |
6. **More models.** For each of the models listed below, predict  $y$  when  $x = 2$ .
- |                                    |                                            |
|------------------------------------|--------------------------------------------|
| a) $\hat{y} = 1.2 + 0.8 \log x$    | d) $\hat{y}^2 = 1.2 + 0.8x$                |
| b) $\log \hat{y} = 1.2 + 0.8x$     | e) $\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$ |
| c) $\ln \hat{y} = 1.2 + 0.8 \ln x$ |                                            |
7. **Gas mileage.** As the example in the chapter indicates, one of the important factors determining a car's *Fuel Efficiency* is its *Weight*. Let's examine this relationship again, for 11 cars.
- a) Describe the association between these variables shown in the scatterplot on the next page.

## RE-EXPRESSION ON THE COMPUTER

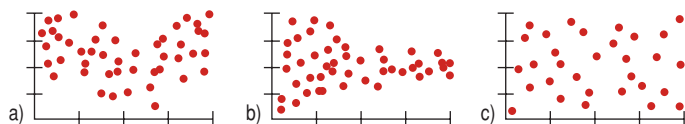
Computers and calculators make it easy to re-express data. Most statistics packages offer a way to re-express and compute with variables. Some packages permit you to specify the power of a re-expression with a slider or other moveable control, possibly while watching the consequences of the re-expression on a plot or analysis. This, of course, is a very effective way to find a good re-expression.

## EXERCISES

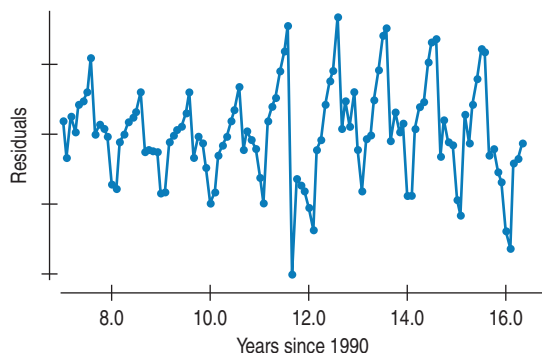
1. **Residuals.** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.



2. **Residuals.** Suppose you have fit a linear model to some data and now take a look at the residuals. For each of the following possible residuals plots, tell whether you would try a re-expression and, if so, why.

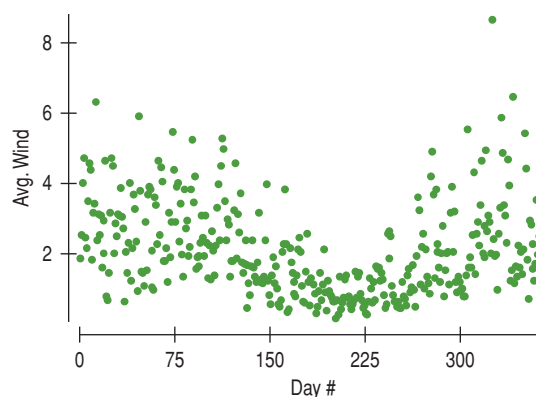


3. **Airline passengers revisited.** In Chapter 9, Exercise 9, we created a linear model describing the trend in the number of passengers departing from the Oakland (CA) airport each month since the start of 1997. Here's the residual plot, but with lines added to show the order of the values in time:

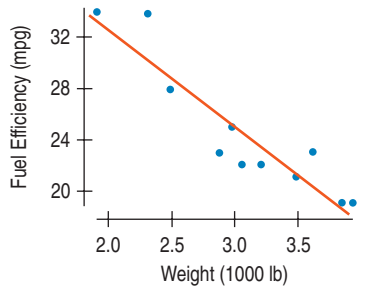


- a) Can you account for the pattern shown here?  
b) Would a re-expression help us deal with this pattern? Explain.

4. **Hopkins winds, revisited.** In Chapter 5, we examined the wind speeds in the Hopkins forest over the course of a year. Here's the scatterplot we saw then:



- a) Describe the pattern you see here.  
b) Should we try re-expressing either variable to make this plot straighter? Explain.
5. **Models.** For each of the models listed below, predict  $y$  when  $x = 2$ .
- |                                     |                                      |
|-------------------------------------|--------------------------------------|
| a) $\ln \hat{y} = 1.2 + 0.8x$       | d) $\hat{y} = 1.2 + 0.8 \ln x$       |
| b) $\sqrt{\hat{y}} = 1.2 + 0.8x$    | e) $\log \hat{y} = 1.2 + 0.8 \log x$ |
| c) $\frac{1}{\hat{y}} = 1.2 + 0.8x$ |                                      |
6. **More models.** For each of the models listed below, predict  $y$  when  $x = 2$ .
- |                                    |                                            |
|------------------------------------|--------------------------------------------|
| a) $\hat{y} = 1.2 + 0.8 \log x$    | d) $\hat{y}^2 = 1.2 + 0.8x$                |
| b) $\log \hat{y} = 1.2 + 0.8x$     | e) $\frac{1}{\sqrt{\hat{y}}} = 1.2 + 0.8x$ |
| c) $\ln \hat{y} = 1.2 + 0.8 \ln x$ |                                            |
7. **Gas mileage.** As the example in the chapter indicates, one of the important factors determining a car's *Fuel Efficiency* is its *Weight*. Let's examine this relationship again, for 11 cars.
- a) Describe the association between these variables shown in the scatterplot on the next page.

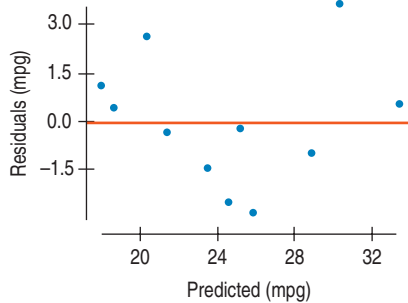


b) Here is the regression analysis for the linear model. What does the slope of the line say about this relationship?

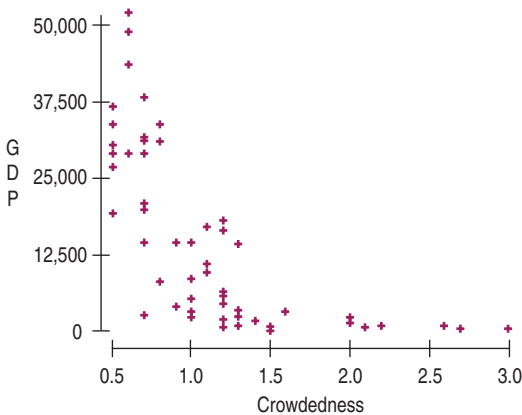
Dependent variable is: Fuel Efficiency  
R-squared = 85.9%

Variable	Coefficient
Intercept	47.9636
Weight	-7.65184

c) Do you think this linear model is appropriate? Use the residuals plot to explain your decision.



**T** 8. **Crowdedness.** In a *Chance* magazine article (Summer 2005), Danielle Vasilescu and Howard Wainer used data from the United Nations Center for Human Settlements to investigate aspects of living conditions for several countries. Among the variables they looked at were the country's per capita gross domestic product (*GDP*, in \$) and *Crowdedness*, defined as the average number of persons per room living in homes there. This scatterplot displays these data for 56 countries:



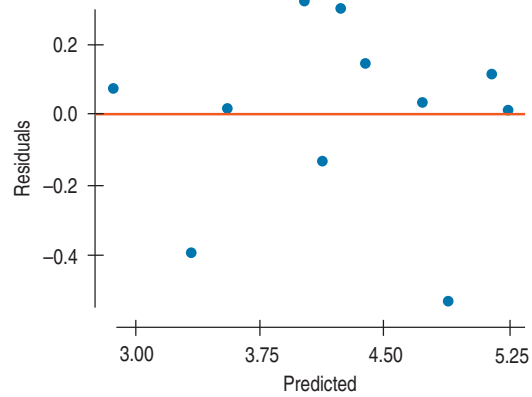
a) Explain why you should re-express these data before trying to fit a model.

b) What re-expression of *GDP* would you try as a starting point?

9. **Gas mileage revisited.** Let's try the re-expressed variable *Fuel Consumption* (gal/100 mi) to examine the fuel efficiency of the 11 cars in Exercise 7. Here are the revised regression analysis and residuals plot:

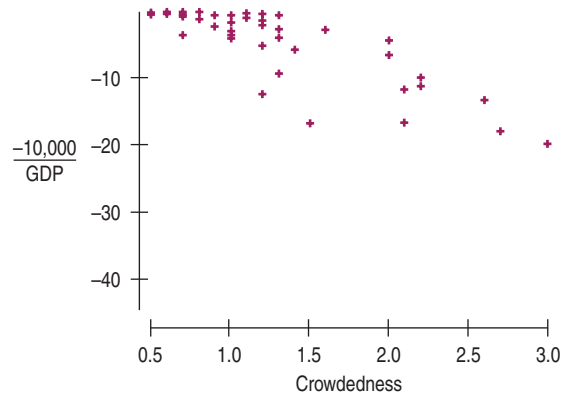
Dependent variable is: Fuel Consumption  
R-squared = 89.2%

Variable	Coefficient
Intercept	0.624932
Weight	1.17791



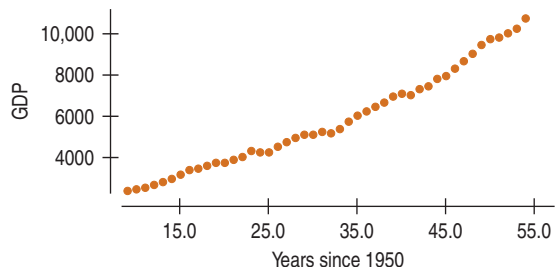
- Explain why this model appears to be better than the linear model.
- Using the regression analysis above, write an equation of this model.
- Interpret the slope of this line.
- Based on this model, how many miles per gallon would you expect a 3500-pound car to get?

10. **Crowdedness again.** In Exercise 8 we looked at United Nations data about a country's *GDP* and the average number of people per room (*Crowdedness*) in housing there. For a re-expression, a student tried the reciprocal  $-10000/GDP$ , representing the number of people per \$10,000 of gross domestic product. Here are the results, plotted against *Crowdedness*:



- Is this a useful re-expression? Explain.
- What re-expression would you suggest this student try next?

11. **GDP.** The scatterplot shows the gross domestic product (GDP) of the United States in billions of dollars plotted against years since 1950.

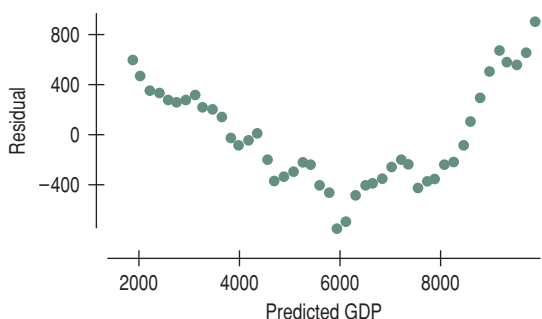


A linear model fit to the relationship looks like this:

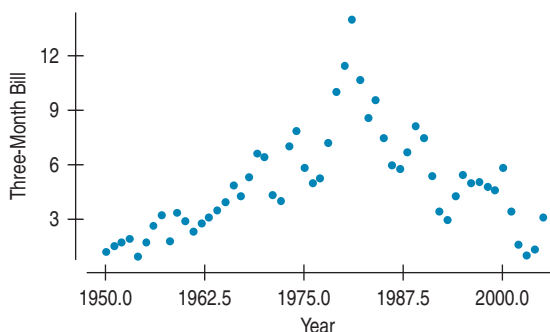
Dependent variable is: GDP  
R-squared = 97.2%    s = 406.6

Variable	Coefficient
Intercept	240.171
Year-1950	177.689

- Does the value 97.2% suggest that this is a good model? Explain.
- Here's a scatterplot of the residuals. Now do you think this is a good model for these data? Explain?



12. **Treasury Bills.** The 3-month Treasury bill interest rate is watched by investors and economists. Here's a scatterplot of the 3-month Treasury bill rate since 1950:

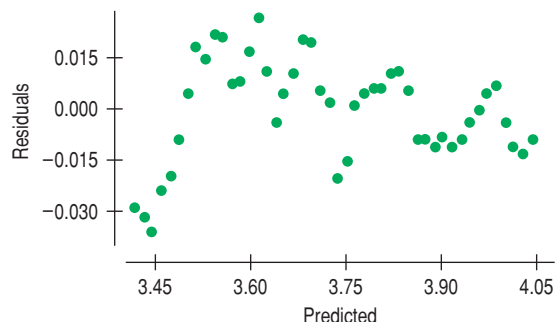


Clearly, the relationship is not linear. Can it be made nearly linear with a re-expression? If so, which one would you suggest? If not, why not?

13. **Better GDP model?** Consider again the post-1950 trend in U.S. GDP we examined in Exercise 11. Here are a regression and residual plot when we use the log of GDP in the model. Is this a better model for GDP? Explain.

Dependent variable is: LogGDP  
R-squared = 99.4%    s = 0.0150

Variable	Coefficient
Intercept	3.29092
Year-1950	0.013881



14. **Pressure.** Scientist Robert Boyle examined the relationship between the volume in which a gas is contained and the pressure in its container. He used a cylindrical container with a moveable top that could be raised or lowered to change the volume. He measured the *Height* in inches by counting equally spaced marks on the cylinder, and measured the *Pressure* in inches of mercury (as in a barometer). Some of his data are listed in the table. Create an appropriate model.

Height	48	44	40	36	32	28
Pressure	29.1	31.9	35.3	39.3	44.2	50.3
Height	24	20	18	16	14	12
Pressure	58.8	70.7	77.9	87.9	100.4	117.6

15. **Brakes.** The table below shows stopping distances in feet for a car tested 3 times at each of 5 speeds. We hope to create a model that predicts *Stopping Distance* from the *Speed* of the car.

Speed (mph)	Stopping Distances (ft)
20	64, 62, 59
30	114, 118, 105
40	153, 171, 165
50	231, 203, 238
60	317, 321, 276

- Explain why a linear model is not appropriate.
  - Re-express the data to straighten the scatterplot.
  - Create an appropriate model.
  - Estimate the stopping distance for a car traveling 55 mph.
  - Estimate the stopping distance for a car traveling 70 mph.
  - How much confidence do you place in these predictions? Why?
16. **Pendulum.** A student experimenting with a pendulum counted the number of full swings the pendulum made in 20 seconds for various lengths of string. Her data are shown on the next page.

Length (in.)	6.5	9	11.5	14.5	18	21	24	27	30	37.5
Number of Swings	22	20	17	16	14	13	13	12	11	10

- Explain why a linear model is not appropriate for using the *Length* of a pendulum to predict the *Number of Swings* in 20 seconds.
- Re-express the data to straighten the scatterplot.
- Create an appropriate model.
- Estimate the number of swings for a pendulum with a 4-inch string.
- Estimate the number of swings for a pendulum with a 48-inch string.
- How much confidence do you place in these predictions? Why?

- T 17. Baseball salaries 2005.** Ballplayers have been signing ever larger contracts. The highest salaries (in millions of dollars per season) for some notable players are given in the following table.

Player	Year	Salary (million \$)
Nolan Ryan	1980	1.0
George Foster	1982	2.0
Kirby Puckett	1990	3.0
Jose Canseco	1990	4.7
Roger Clemens	1991	5.3
Ken Griffey, Jr.	1996	8.5
Albert Belle	1997	11.0
Pedro Martinez	1998	12.5
Mike Piazza	1999	12.5
Mo Vaughn	1999	13.3
Kevin Brown	1999	15.0
Carlos Delgado	2001	17.0
Alex Rodriguez	2001	22.0
Manny Ramirez	2004	22.5
Alex Rodriguez	2005	26.0

- Examine a scatterplot of the data. Does it look straight?
- Find the regression of *Salary* vs. *Year* and plot the residuals. Do they look straight?
- Re-express the data, if necessary, to straighten the relationship.
- What model would you report for the trend in salaries?

- T 18. Planet distances and years 2006.** At a meeting of the International Astronomical Union (IAU) in Prague in 2006, Pluto was determined not to be a planet, but rather the largest member of the Kuiper belt of icy objects. Let's examine some facts. Here is a table of the 9 sun-orbiting objects formerly known as planets:

Planet	Position Number	Distance from Sun (million miles)	Length of Year (Earth years)
Mercury	1	36	0.24
Venus	2	67	0.61
Earth	3	93	1.00
Mars	4	142	1.88
Jupiter	5	484	11.86
Saturn	6	887	29.46
Uranus	7	1784	84.07
Neptune	8	2796	164.82
Pluto	9	3707	247.68

- Plot the *Length* of the year against the *Distance* from the sun. Describe the shape of your plot.
- Re-express one or both variables to straighten the plot. Use the re-expressed data to create a model describing the length of a planet's year based on its distance from the sun.
- Comment on how well your model fits the data.

- T 19. Planet distances and order 2006.** Let's look again at the pattern in the locations of the planets in our solar system seen in the table in Exercise 18.

- Re-express the distances to create a model for the *Distance* from the sun based on the planet's *Position*.
- Based on this model, would you agree with the International Astronomical Union that Pluto is not a planet? Explain.

- T 20. Planets 2006, part 3.** The asteroid belt between Mars and Jupiter may be the remnants of a failed planet. If so, then Jupiter is really in position 6, Saturn is in 7, and so on. Repeat Exercise 19, using this revised method of numbering the positions. Which method seems to work better?

- T 21. Eris: Planets 2006, part 4.** In July 2005, astronomers Mike Brown, Chad Trujillo, and David Rabinowitz announced the discovery of a sun-orbiting object, since named Eris,<sup>6</sup> that is 5% larger than Pluto. Eris orbits the sun once every 560 earth years at an average distance of about 6300 million miles from the sun. Based on its *Position*, how does Eris's *Distance* from the sun (re-expressed to logs) compare with the prediction made by your model of Exercise 19?

- T 22. Models and laws: Planets 2006 part 5.** The model you found in Exercise 18 is a relationship noted in the 17th century by Kepler as his Third Law of Planetary Motion. It was subsequently explained as a consequence of Newton's Law of Gravitation. The models

<sup>6</sup>Eris is the Greek goddess of warfare and strife who caused a quarrel among the other goddesses that led to the Trojan war. In the astronomical world, Eris stirred up trouble when the question of its proper designation led to the raucous meeting of the IAU in Prague where IAU members voted to demote Pluto and Eris to dwarf-planet status—<http://www.gps.caltech.edu/~mbrown/planetlila/#paper>.

for Exercises 19–21 relate to what is sometimes called the Titius-Bode “law,” a pattern noticed in the 18th century but lacking any scientific explanation.

Compare how well the re-expressed data are described by their respective linear models. What aspect of the model of Exercise 18 suggests that we have found a physical law? In the future, we may learn enough about a planetary system around another star to tell whether the Titius-Bode pattern applies there. If you discovered that another planetary system followed the same pattern, how would it change your opinion about whether this is a real natural “law”? What would you think if the next system we find does not follow this pattern?

23. **Logs (not logarithms).** The value of a log is based on the number of board feet of lumber the log may contain. (A board foot is the equivalent of a piece of wood 1 inch thick, 12 inches wide, and 1 foot long. For example, a  $2'' \times 4''$  piece that is 12 feet long contains 8 board feet.) To estimate the amount of lumber in a log, buyers measure the diameter inside the bark at the smaller end. Then they look in a table based on the Doyle Log Scale. The table below shows the estimates for logs 16 feet long.

Diameter of Log	8"	12"	16"	20"	24"	28"
Board Feet	16	64	144	256	400	576

- a) What model does this scale use?  
 b) How much lumber would you estimate that a log 10 inches in diameter contains?  
 c) What does this model suggest about logs 36 inches in diameter?
- T 24. **Weightlifting 2004.** Listed below are the gold medal-winning men’s weight-lifting performances at the 2004 Olympics.

Weight Class (kg)	Winner (country)	Weight Lifted (kg)
56	Halil Mutlu (Turkey)	295.0
62	Zhiyong Shi (China)	325.0
69	Guozheng Zhang (China)	347.5
77	Taner Sagir (Turkey)	375.0
85	George Asanidze (Georgia)	382.5
94	Milen Dobrev (Bulgaria)	407.5
105	Dmitry Berestov (Russia)	425.0

- a) Create a linear model for the *Weight Lifted* in each *Weight Class*.  
 b) Check the residuals plot. Is your linear model appropriate?  
 c) Create a better model.  
 d) Explain why you think your new model is better.  
 e) Based on your model, which of the medalists turned in the most surprising performance? Explain.
- T 25. **Life expectancy.** The data in the next column list the *Life Expectancy* for white males in the United States every decade during the last century (1 = 1900 to 1910, 2 = 1911

to 1920, etc.). Create a model to predict future increases in life expectancy. (National Vital Statistics Report)

Decade	1	2	3	4	5	6	7	8	9	10
Life exp.	48.6	54.4	59.7	62.1	66.5	67.4	68.0	70.7	72.7	74.9

- T 26. **Lifting more weight 2004.** In Exercise 24 you examined the winning weight-lifting performances for the 2004 Olympics. One of the competitors turned in a performance that appears not to fit the model you created.
- a) Consider that competitor to be an outlier. Eliminate that data point and re-create your model.  
 b) Using this revised model, how much would you have expected the outlier competitor to lift?  
 c) Explain the meaning of the residual from your new model for that competitor.
- T 27. **Slower is cheaper?** Researchers studying how a car’s *Fuel Efficiency* varies with its *Speed* drove a compact car 200 miles at various speeds on a test track. Their data are shown in the table.

Speed (mph)	35	40	45	50	55	60	65	70	75
Fuel Eff. (mpg)	25.9	27.7	28.5	29.5	29.2	27.4	26.4	24.2	22.8

Create a linear model for this relationship and report any concerns you may have about the model.

- T 28. **Orange production.** The table below shows that as the number of oranges on a tree increases, the fruit tends to get smaller. Create a model for this relationship, and express any concerns you may have.

Number of Oranges/Tree	Average Weight/Fruit (lb)
50	0.60
100	0.58
150	0.56
200	0.55
250	0.53
300	0.52
350	0.50
400	0.49
450	0.48
500	0.46
600	0.44
700	0.42
800	0.40
900	0.38

- T 29. **Years to live 2003.** Insurance companies and other organizations use actuarial tables to estimate the remaining lifespans of their customers. On the next page are the estimated additional years of life for black males in the United States, according to a 2003 National Vital Statistics Report. ([www.cdc.gov/nchs/deaths.htm](http://www.cdc.gov/nchs/deaths.htm))



Age	10	20	30	40	50	60	70	80	90	100
Years Left	60.3	50.7	41.8	32.9	24.8	17.9	12.1	7.9	5.0	3.0

- Find a re-expression to create an appropriate model.
- Predict the lifespan of an 18-year-old black man.
- Are you satisfied that your model has accounted for the relationship between *Years Left* and *Age*? Explain.

- T 30. Tree growth.** A 1996 study examined the growth of grapefruit trees in Texas, determining the average trunk *Diameter* (in inches) for trees of varying *Ages*:

Age (yr)	2	4	6	8	10	12	14	16	18	20
Diameter (in.)	2.1	3.9	5.2	6.2	6.9	7.6	8.3	9.1	10.0	11.4

- Fit a linear model to these data. What concerns do you have about the model?
- If data had been given for individual trees instead of averages, would you expect the fit to be stronger, less strong, or about the same? Explain.



## JUST CHECKING Answers

- Counts are often best transformed by using the square root.
- None. The relationship is already straight.
- Even though, technically, the population values are counts, you should probably try a stronger transformation like  $\log(\text{population})$  because populations grow in proportion to their size.

## PART

## REVIEW OF PART II

### Exploring Relationships Between Variables

#### Quick Review

You have now survived your second major unit of Statistics. Here's a brief summary of the key concepts and skills:

- ▶ We treat data two ways: as categorical and as quantitative.
- ▶ To explore relationships in categorical data, check out Chapter 3.
- ▶ To explore relationships in quantitative data:
  - Make a picture. Use a scatterplot. Put the explanatory variable on the  $x$ -axis and the response variable on the  $y$ -axis.
  - Describe the association between two quantitative variables in terms of direction, form, and strength.
  - The amount of scatter determines the strength of the association.
  - If, as one variable increases so does the other, the association is positive. If one increases as the other decreases, it's negative.
  - If the form of the association is linear, calculate a correlation to measure its strength numerically, and do a regression analysis to model it.
  - Correlations closer to  $-1$  or  $+1$  indicate stronger linear associations. Correlations near  $0$  indicate weak linear relationships, but other forms of association may still be present.
  - The line of best fit is also called the least squares regression line because it minimizes the sum of the squared residuals.
  - The regression line predicts values of the response variable from values of the explanatory variable.

- A residual is the difference between the true value of the response variable and the value predicted by the regression model.
- The slope of the line is a rate of change, best described in " $y$ -units" per " $x$ -unit."
- $R^2$  gives the fraction of the variation in the response variable that is accounted for by the model.
- The standard deviation of the residuals measures the amount of scatter around the line.
- Outliers and influential points can distort any of our models.
- If you see a pattern (a curve) in the residuals plot, your chosen model is not appropriate; use a different model. You may, for example, straighten the relationship by re-expressing one of the variables.
- To straighten bent relationships, re-express the data using logarithms or a power (squares, square roots, reciprocals, etc.).
- Always remember that an association is not necessarily an indication that one of the variables causes the other.

Need more help with some of this? Try rereading some sections of Chapters 7 through 10. And go on to the next page for more opportunities to review these concepts and skills.

*"One must learn by doing the thing; though you think you know it, you have no certainty until you try."*

—Sophocles (495–406 BCE)

Age	10	20	30	40	50	60	70	80	90	100
Years Left	60.3	50.7	41.8	32.9	24.8	17.9	12.1	7.9	5.0	3.0

- Find a re-expression to create an appropriate model.
- Predict the lifespan of an 18-year-old black man.
- Are you satisfied that your model has accounted for the relationship between *Years Left* and *Age*? Explain.

- T 30. Tree growth.** A 1996 study examined the growth of grapefruit trees in Texas, determining the average trunk *Diameter* (in inches) for trees of varying *Ages*:

Age (yr)	2	4	6	8	10	12	14	16	18	20
Diameter (in.)	2.1	3.9	5.2	6.2	6.9	7.6	8.3	9.1	10.0	11.4

- Fit a linear model to these data. What concerns do you have about the model?
- If data had been given for individual trees instead of averages, would you expect the fit to be stronger, less strong, or about the same? Explain.



## JUST CHECKING Answers

- Counts are often best transformed by using the square root.
- None. The relationship is already straight.
- Even though, technically, the population values are counts, you should probably try a stronger transformation like  $\log(\text{population})$  because populations grow in proportion to their size.

## PART



## REVIEW OF PART II

### Exploring Relationships Between Variables

#### Quick Review

You have now survived your second major unit of Statistics. Here's a brief summary of the key concepts and skills:

- ▶ We treat data two ways: as categorical and as quantitative.
- ▶ To explore relationships in categorical data, check out Chapter 3.
- ▶ To explore relationships in quantitative data:
  - Make a picture. Use a scatterplot. Put the explanatory variable on the  $x$ -axis and the response variable on the  $y$ -axis.
  - Describe the association between two quantitative variables in terms of direction, form, and strength.
  - The amount of scatter determines the strength of the association.
  - If, as one variable increases so does the other, the association is positive. If one increases as the other decreases, it's negative.
  - If the form of the association is linear, calculate a correlation to measure its strength numerically, and do a regression analysis to model it.
  - Correlations closer to  $-1$  or  $+1$  indicate stronger linear associations. Correlations near  $0$  indicate weak linear relationships, but other forms of association may still be present.
  - The line of best fit is also called the least squares regression line because it minimizes the sum of the squared residuals.
  - The regression line predicts values of the response variable from values of the explanatory variable.

- A residual is the difference between the true value of the response variable and the value predicted by the regression model.
- The slope of the line is a rate of change, best described in " $y$ -units" per " $x$ -unit."
- $R^2$  gives the fraction of the variation in the response variable that is accounted for by the model.
- The standard deviation of the residuals measures the amount of scatter around the line.
- Outliers and influential points can distort any of our models.
- If you see a pattern (a curve) in the residuals plot, your chosen model is not appropriate; use a different model. You may, for example, straighten the relationship by re-expressing one of the variables.
- To straighten bent relationships, re-express the data using logarithms or a power (squares, square roots, reciprocals, etc.).
- Always remember that an association is not necessarily an indication that one of the variables causes the other.

Need more help with some of this? Try rereading some sections of Chapters 7 through 10. And go on to the next page for more opportunities to review these concepts and skills.

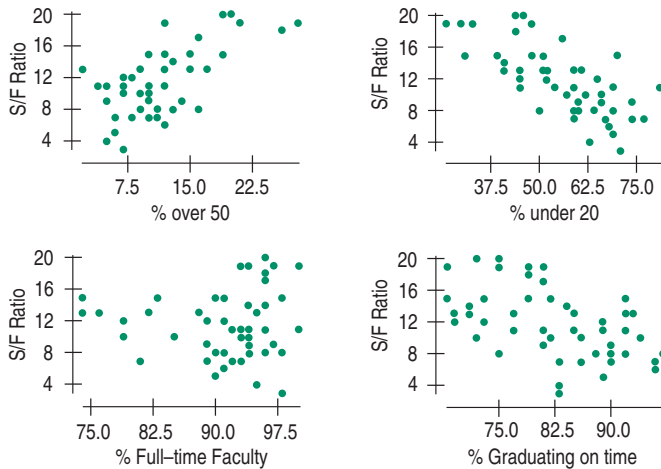
*"One must learn by doing the thing; though you think you know it, you have no certainty until you try."*

—Sophocles (495–406 BCE)

## REVIEW EXERCISES

1. **College.** Every year *US News and World Report* publishes a special issue on many U.S. colleges and universities. The scatterplots below have *Student/Faculty Ratio* (number of students per faculty member) for the colleges and universities on the *y*-axes plotted against 4 other variables. The correct correlations for these scatterplots appear in this list. Match them.

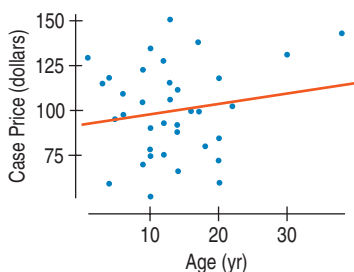
− 0.98 − 0.71 − 0.51 0.09 0.23 0.69



2. **Togetherness.** Are good grades in high school associated with family togetherness? A random sample of 142 high school students was asked how many meals per week their families ate together. Their responses produced a mean of 3.78 meals per week, with a standard deviation of 2.2. Researchers then matched these responses against the students' grade point averages (GPAs). The scatterplot appeared to be reasonably linear, so they created a line of regression. No apparent pattern emerged in the residuals plot. The equation of the line was  $GPA = 2.73 + 0.11 \text{ Meals}$ .

- Interpret the *y*-intercept in this context.
- Interpret the slope in this context.
- What was the mean GPA for these students?
- If a student in this study had a negative residual, what did that mean?
- Upon hearing of this study, a counselor recommended that parents who want to improve the grades their children get should get the family to eat together more often. Do you agree with this interpretation? Explain.

3. **Vineyards.** Here are the scatterplot and regression analysis for *Case Prices* of 36 wines from vineyards in the Finger Lakes region of New York State and the *Ages* of the vineyards.



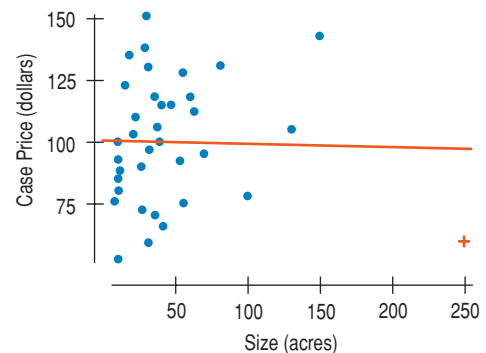
Dependent variable is: Case Price

R-squared = 2.7%

Variable	Coefficient
Constant	92.7650
Age	0.567284

- Does it appear that vineyards in business longer get higher prices for their wines? Explain.
- What does this analysis tell us about vineyards in the rest of the world?
- Write the regression equation.
- Explain why that equation is essentially useless.

4. **Vineyards again.** Instead of *Age*, perhaps the *Size* of the vineyard (in acres) is associated with the price of the wines. Look at the scatterplot:



- Do you see any evidence of an association?
- What concern do you have about this scatterplot?
- If the red "+" data point is removed, would the correlation become stronger or weaker? Explain.
- If the red "+" data point is removed, would the slope of the line increase or decrease? Explain.

5. **More twins 2004?** As the table shows, the number of twins born in the United States has been increasing. ([www.cdc.gov/nchs/births.htm](http://www.cdc.gov/nchs/births.htm))

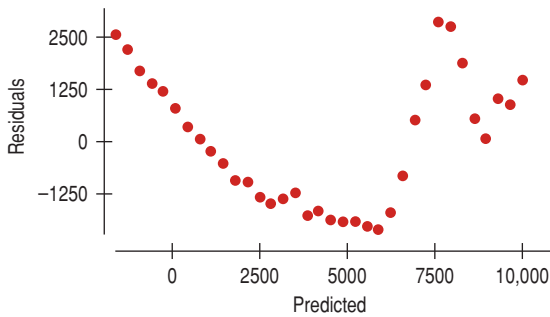
Year	Twin Births	Year	Twin Births
1980	68,339	1993	96,445
1981	70,049	1994	97,064
1982	71,631	1995	96,736
1983	72,287	1996	100,750
1984	72,949	1997	104,137
1985	77,102	1998	110,670
1986	79,485	1999	114,307
1987	81,778	2000	118,916
1988	85,315	2001	121,246
1989	90,118	2002	125,134
1990	93,865	2003	128,665
1991	94,779	2004	132,219
1992	95,372		

- Find the equation of the regression line for predicting the number of twin births.
  - Explain in this context what the slope means.
  - Predict the number of twin births in the United States for the year 2010. Comment on your faith in that prediction.
  - Comment on the residuals plot.
6. **Dow Jones 2006.** The Dow Jones stock index measures the performance of the stocks of America's largest companies (<http://finance.yahoo.com>). A regression of the Dow prices on years 1972–2006 looks like this:

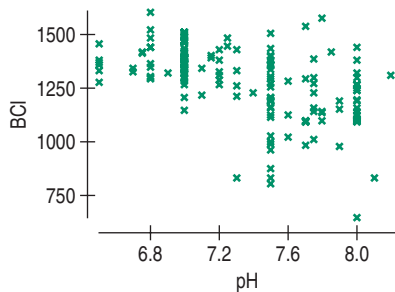
Dependent variable is: Dow Index  
 R-squared = 83.5%     $s = 1577$

Variable	Coefficient
Intercept	-2294.01
Year since 1970	341.095

- What is the correlation between *Dow Index* and *Year*?
- Write the regression equation.
- Explain in this context what the equation says.
- Here's a scatterplot of the residuals. Which assumption(s) of the regression analysis appear to be violated?



7. **Acid rain.** Biologists studying the effects of acid rain on wildlife collected data from 163 streams in the Adirondack Mountains. They recorded the *pH* (acidity) of the water and the *BCI*, a measure of biological diversity, and they calculated  $R^2 = 27\%$ . Here's a scatterplot of *BCI* against *pH*:



- What is the correlation between *pH* and *BCI*?
  - Describe the association between these two variables.
  - If a stream has average *pH*, what would you predict about the *BCI*?
  - In a stream where the *pH* is 3 standard deviations above average, what would you predict about the *BCI*?
8. **Manatees 2005.** Marine biologists warn that the growing number of powerboats registered in Florida threatens the existence of manatees. The data below come from the

Florida Fish and Wildlife Conservation Commission ([www.floridamarine.org](http://www.floridamarine.org)) and the National Marine Manufacturers Association ([www.nmma.org/facts](http://www.nmma.org/facts)).

Year	Manatees Killed	Powerboat Registrations (in 1000s)
1982	13	447
1983	21	460
1984	24	481
1985	16	498
1986	24	513
1987	20	512
1988	15	527
1989	34	559
1990	33	585
1992	33	614
1993	39	646
1994	43	675
1995	50	711
1996	47	719
1997	53	716
1998	38	716
1999	35	716
2000	49	735
2001	81	860
2002	95	923
2003	73	940
2004	69	946
2005	79	974

- In this context, which is the explanatory variable?
  - Make a scatterplot of these data and describe the association you see.
  - Find the correlation between *Boat Registrations* and *Manatee Deaths*.
  - Interpret the value of  $R^2$ .
  - Does your analysis prove that powerboats are killing manatees?
9. **A manatee model 2005.** Continue your analysis of the manatee situation from the previous exercise.
- Create a linear model of the association between *Manatee Deaths* and *Powerboat Registrations*.
  - Interpret the slope of your model.
  - Interpret the *y*-intercept of your model.
  - How accurately did your model predict the high number of manatee deaths in 2005?
  - Which is better for the manatees, positive residuals or negative residuals? Explain.
  - What does your model suggest about the future for the manatee?
10. **Grades.** A Statistics instructor created a linear regression equation to predict students' final exam scores from their midterm exam scores. The regression equation was  $\widehat{Fin} = 10 + 0.9 Mid$ .
- If Susan scored a 70 on the midterm, what did the instructor predict for her score on the final?

- b) Susan got an 80 on the final. How big is her residual?
- c) If the standard deviation of the final was 12 points and the standard deviation of the midterm was 10 points, what is the correlation between the two tests?
- d) How many points would someone need to score on the midterm to have a predicted final score of 100?
- e) Suppose someone scored 100 on the final. Explain why you can't estimate this student's midterm score from the information given.
- f) One of the students in the class scored 100 on the midterm but got overconfident, slacked off, and scored only 15 on the final exam. What is the residual for this student?
- g) No other student in the class "achieved" such a dramatic turnaround. If the instructor decides not to include this student's scores when constructing a new regression model, will the  $R^2$  value of the regression increase, decrease, or remain the same? Explain.
- h) Will the slope of the new line increase or decrease?

**T** 11. **Traffic.** Highway planners investigated the relationship between traffic *Density* (number of automobiles per mile) and the average *Speed* of the traffic on a moderately large city thoroughfare. The data were collected at the same location at 10 different times over a span of 3 months. They found a mean traffic *Density* of 68.6 cars per mile (cpm) with standard deviation of 27.07 cpm. Overall, the cars' average *Speed* was 26.38 mph, with standard deviation of 9.68 mph. These researchers found the regression line for these data to be  $\widehat{Speed} = 50.55 - 0.352 \text{ Density}$ .

- a) What is the value of the correlation coefficient between *Speed* and *Density*?
- b) What percent of the variation in average *Speed* is explained by traffic *Density*?
- c) Predict the average *Speed* of traffic on the thoroughfare when the traffic *Density* is 50 cpm.
- d) What is the value of the residual for a traffic *Density* of 56 cpm with an observed *Speed* of 32.5 mph?
- e) The data set initially included the point  $\text{Density} = 125$  cpm,  $\text{Speed} = 55$  mph. This point was considered an outlier and was not included in the analysis. Will the slope increase, decrease, or remain the same if we redo the analysis and include this point?
- f) Will the correlation become stronger, weaker, or remain the same if we redo the analysis and include this point (125,55)?
- g) A European member of the research team measured the *Speed* of the cars in kilometers per hour (1 km  $\approx$  0.62 miles) and the traffic *Density* in cars per kilometer. Find the value of his calculated correlation between speed and density.

**T** 12. **Cramming.** One Thursday, researchers gave students enrolled in a section of basic Spanish a set of 50 new vocabulary words to memorize. On Friday the students took a vocabulary test. When they returned to class the following Monday, they were retested—without advance warning. Here are the test scores for the 25 students.

Fri.	Mon.	Fri.	Mon.	Fri.	Mon.
42	36	48	37	39	41
44	44	43	41	46	32
45	46	45	32	37	36
48	38	47	44	40	31
44	40	50	47	41	32
43	38	34	34	48	39
41	37	38	31	37	31
35	31	43	40	36	41
43	32				

- a) What is the correlation between *Friday* and *Monday* scores?
- b) What does a scatterplot show about the association between the scores?
- c) What does it mean for a student to have a positive residual?
- d) What would you predict about a student whose *Friday* score was one standard deviation below average?
- e) Write the equation of the regression line.
- f) Predict the *Monday* score of a student who earned a 40 on Friday.

13. **Correlations.** What factor most explains differences in *Fuel Efficiency* among cars? Below is a correlation matrix exploring that relationship for the car's *Weight*, *Horsepower*, engine size (*Displacement*), and number of *Cylinders*.

	MPG	Weight	Horsepower	Displacement	Cylinders
MPG	1.000				
Weight	-0.903	1.000			
Horsepower	-0.871	0.917	1.000		
Displacement	-0.786	0.951	0.872	1.000	
Cylinders	-0.806	0.917	0.864	0.940	1.000

- a) Which factor seems most strongly associated with *Fuel Efficiency*?
- b) What does the negative correlation indicate?
- c) Explain the meaning of  $R^2$  for that relationship.

**T** 14. **Autos revisited.** Look again at the correlation table for cars in the previous exercise.

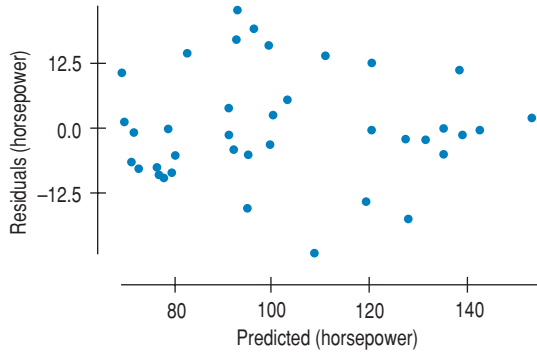
- a) Which two variables in the table exhibit the strongest association?
- b) Is that strong association necessarily cause-and-effect? Offer at least two explanations why that association might be so strong.
- c) Engine displacements for U.S.-made cars are often measured in cubic inches. For many foreign cars, the units are either cubic centimeters or liters. How would changing from cubic inches to liters affect the calculated correlations involving *Displacement*?
- d) What would you predict about the *Fuel Efficiency* of a car whose engine *Displacement* is one standard deviation above the mean?

**T** 15. **Cars, one more time!** Can we predict the *Horsepower* of the engine that manufacturers will put in a car by

knowing the *Weight* of the car? Here are the regression analysis and residuals plot:

Dependent variable is: Horsepower  
R-squared = 84.1%

Variable	Coefficient
Intercept	3.49834
Weight	34.3144

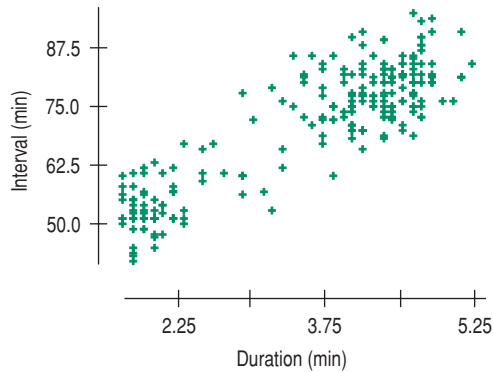


- Write the equation of the regression line.
- Do you think the car's *Weight* is measured in pounds or thousands of pounds? Explain.
- Do you think this linear model is appropriate? Explain.
- The highest point in the residuals plot, representing a residual of 22.5 horsepower, is for a Chevy weighing 2595 pounds. How much horsepower does this car have?

16. **Colorblind.** Although some women are colorblind, this condition is found primarily in men. Why is it wrong to say there's a strong correlation between *Sex* and *Colorblindness*?

T 17. **Old Faithful.** There is evidence that eruptions of Old Faithful can best be predicted by knowing the duration of the previous eruption.

- Describe what you see in the scatterplot of *Intervals* between eruptions vs. *Duration* of the previous eruption.



- Write the equation of the line of best fit. Here's the regression analysis:

Dependent variable is: Interval  
R-squared = 77.0%

Variable	Coefficient
Intercept	33.9668
Duration	10.3582

- Carefully explain what the slope of the line means in this context.
- How accurate do you expect predictions based on this model to be? Cite statistical evidence.
- If you just witnessed an eruption that lasted 4 minutes, how long do you predict you'll have to wait to see the next eruption?
- So you waited, and the next eruption came in 79 minutes. Use this as an example to define a residual.

T 18. **Which croc?** The ranges inhabited by the Indian gharial crocodile and the Australian saltwater crocodile overlap in Bangladesh. Suppose a very large crocodile skeleton is found there, and we wish to determine the species of the animal. Wildlife scientists have measured the lengths of the heads and the complete bodies of several crocs (in centimeters) of each species, creating the regression analyses below:

**Indian Crocodile**

Dependent variable is: IBody  
R-squared = 97.2%

Variable	Coefficient
Intercept	-69.3693
IHead	7.40004

**Australian Crocodile**

Dependent variable is: ABody  
R-squared = 98.0%

Variable	Coefficient
Intercept	-20.2245
AHead	7.71726

- Do the associations between the sizes of the heads and bodies of the two species appear to be strong? Explain.
- In what ways are the two relationships similar? Explain.
- What is different about the two models? What does that mean?
- The crocodile skeleton found had a head length of 62 cm and a body length of 380 cm. Which species do you think it was? Explain why.

T 19. **How old is that tree?** One can determine how old a tree is by counting its rings, but that requires cutting the tree down. Can we estimate the tree's age simply from its diameter? A forester measured 27 trees of the same species that had been cut down, and counted the rings to determine the ages of the trees.

Diameter (in.)	Age (yr)	Diameter (in.)	Age (yr)
1.8	4	10.3	23
1.8	5	14.3	25
2.2	8	13.2	28
4.4	8	9.9	29
6.6	8	13.2	30
4.4	10	15.4	30
7.7	10	17.6	33
10.8	12	14.3	34
7.7	13	15.4	35
5.5	14	11.0	38
9.9	16	15.4	38
10.1	18	16.5	40
12.1	20	16.5	42
12.8	22		

- a) Find the correlation between *Diameter* and *Age*. Does this suggest that a linear model may be appropriate? Explain.
- b) Create a scatterplot and describe the association.
- c) Create the linear model.
- d) Check the residuals. Explain why a linear model is probably not appropriate.
- e) If you used this model, would it generally overestimate or underestimate the ages of very large trees? Explain.

**T 20. Improving trees.** In the last exercise you saw that the linear model had some deficiencies. Let's create a better model.

- a) Perhaps the cross-sectional area of a tree would be a better predictor of its age. Since area is measured in square units, try re-expressing the data by squaring the diameters. Does the scatterplot look better?
- b) Create a model that predicts *Age* from the square of the *Diameter*.
- c) Check the residuals plot for this new model. Is this model more appropriate? Why?
- d) Estimate the age of a tree 18 inches in diameter.

**21. New homes.** A real estate agent collects data to develop a model that will use the *Size* of a new home (in square feet) to predict its *Sale Price* (in thousands of dollars). Which of these is most likely to be the slope of the regression line: 0.008, 0.08, 0.8, or 8? Explain.

**T 22. Smoking and pregnancy 2003.** The organization Kids Count monitors issues related to children. The table shows a 50-state average of the percent of expectant mothers who smoked cigarettes during their pregnancies.

Year	% Smoking While Pregnant	Year	% Smoking While Pregnant
1990	19.2	1997	14.9
1991	18.7	1998	14.8
1992	17.9	1999	14.1
1993	16.8	2000	14.0
1994	16.0	2001	13.8
1995	15.4	2002	13.3
1996	15.3	2003	12.7

- a) Create a scatterplot and describe the trend you see.
- b) Find the correlation.
- c) How is the value of the correlation affected by the fact that the data are averages rather than percentages for each of the 50 states?
- d) Write a linear model and interpret the slope in context.

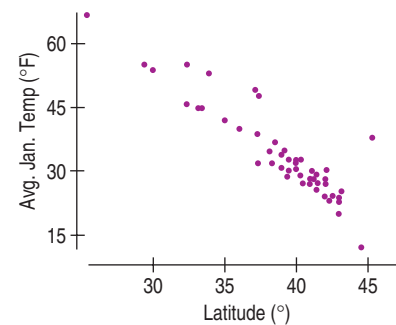
**T 23. No smoking?** The downward trend in smoking you saw in the last exercise is good news for the health of babies, but will it ever stop?

- a) Explain why you can't use the linear model you created in Exercise 22 to see when smoking during pregnancy will cease altogether.
- b) Create a model that could estimate the year in which the level of smoking would be 0%.
- c) Comment on the reliability of such a prediction.

**24. Tips.** It's commonly believed that people use tips to reward good service. A researcher for the hospitality industry examined tips and ratings of service quality from 2645 dining parties at 21 different restaurants. The correlation between ratings of service and tip percentages was 0.11. (M. Lynn and M. McCall, "Gratitude and Gratitude." *Journal of Socio-Economics* 29: 203–214)

- a) Describe the relationship between *Quality of Service* and *Tip Size*.
- b) Find and interpret the value of  $R^2$  in this context.

**T 25. US Cities.** Data from 50 large U.S. cities show the mean *January Temperature* and the *Latitude*. Describe what you see in the scatterplot.



**T 26. Correlations.** The study of U.S. cities in Exercise 25 found the mean *January Temperature* (degrees Fahrenheit), *Altitude* (feet above sea level), and *Latitude* (degrees north of the equator) for 55 cities. Here's the correlation matrix:

	Jan. Temp	Latitude	Altitude
Jan. Temp	1.000		
Latitude	-0.848	1.000	
Altitude	-0.369	0.184	1.000

- a) Which seems to be more useful in predicting *January Temperature*—*Altitude* or *Latitude*? Explain.
- b) If the *Temperature* were measured in degrees Celsius, what would be the correlation between *Temperature* and *Latitude*?
- c) If the *Temperature* were measured in degrees Celsius and the *Altitude* in meters, what would be the correlation? Explain.
- d) What would you predict about the *January Temperatures* in a city whose *Altitude* is two standard deviations higher than the average *Altitude*?

**T 27. Winter in the city.** Summary statistics for the data relating the latitude and average January temperature for 55 large U.S. cities are given below.

Variable	Mean	StdDev
Latitude	39.02	5.42
JanTemp	26.44	13.49

Correlation = -0.848

- a) What percent of the variation in *January Temperatures* can be explained by variation in *Latitude*?
- b) What is indicated by the fact that the correlation is negative?
- c) Write the equation of the line of regression for predicting *January Temperature* from *Latitude*.
- d) Explain what the slope of the line means.

- e) Do you think the  $y$ -intercept is meaningful? Explain.
- f) The latitude of Denver is  $40^\circ$  N. Predict the mean January temperature there.
- g) What does it mean if the residual for a city is positive?

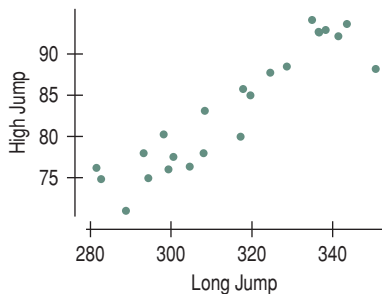
**28. Depression.** The September 1998 issue of the *American Psychologist* published an article by Kraut et al. that reported on an experiment examining “the social and psychological impact of the Internet on 169 people in 73 households during their first 1 to 2 years online.” In the experiment, 73 households were offered free Internet access for 1 or 2 years in return for allowing their time and activity online to be tracked. The members of the households who participated in the study were also given a battery of tests at the beginning and again at the end of the study. The conclusion of the study made news headlines: Those who spent more time online tended to be more depressed at the end of the experiment. Although the paper reports a more complex model, the basic result can be summarized in the following regression of *Depression* (at the end of the study, in “depression scale units”) vs. *Internet Use* (in mean hours per week):

Dependent variable is: Depression  
 R-squared = 4.6%  
 $s = 0.4563$

Variable	Coefficient
Intercept	0.5655
Internet use	0.0199

The news reports about this study clearly concluded that using the Internet causes depression. Discuss whether such a conclusion can be drawn from this regression. If so, discuss the supporting evidence. If not, say why not.

**T 29. Jumps 2004.** How are Olympic performances in various events related? The plot shows winning long-jump and high-jump distances, in inches, for the Summer Olympics from 1912 through 2004.



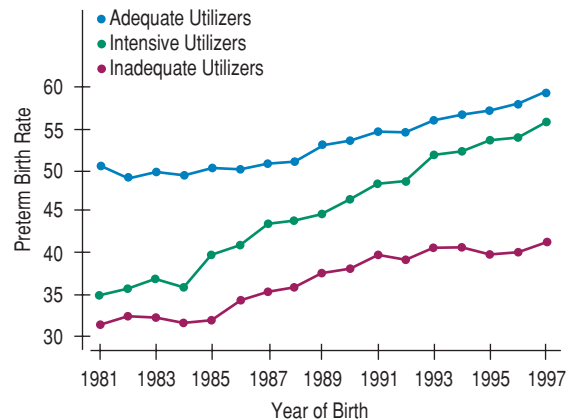
- a) Describe the association.
- b) Do long-jump performances somehow influence the high-jumpers? How do you account for the relationship you see?
- c) The correlation for the given scatterplot is 0.925, but at the Olympics these jumps are actually measured in meters rather than inches. Does that make the actual correlation higher or lower?
- d) What would you predict about the long jump in a year when the high-jumper jumped one standard deviation better than the average high jump?

**T 30. Modeling jumps.** Here are the summary statistics for the Olympic long jumps and high jumps displayed in the scatterplot above:

Event	Mean	StdDev
Long Jump	316.04	20.85
High Jump	83.85	7.46

Correlation = 0.925

- a) Write the equation of the line of regression for estimating *High Jump* from *Long Jump*.
  - b) Interpret the slope of the line.
  - c) In a year when the long jump is 350 inches, what high jump would you predict?
  - d) Why can't you use this line to estimate the long jump for a year when you know the high jump was 85 inches?
  - e) Write the equation of the line you need to make that prediction.
- 31. French.** Consider the association between a student's score on a French vocabulary test and the weight of the student. What direction and strength of correlation would you expect in each of the following situations? Explain.
- a) The students are all in third grade.
  - b) The students are in third through twelfth grades in the same school district.
  - c) The students are in tenth grade in France.
  - d) The students are in third through twelfth grades in France.
- 32. Twins.** Twins are often born after a pregnancy that lasts less than 9 months. The graph from the *Journal of the American Medical Association (JAMA)* shows the rate of preterm twin births in the United States over the past 20 years. In this study, *JAMA* categorized mothers by the level of prenatal medical care they received: inadequate, adequate, or intensive.
- a) Describe the overall trend in preterm twin births.
  - b) Describe any differences you see in this trend, depending on the level of prenatal medical care the mother received.
  - c) Should expectant mothers be advised to cut back on the level of medical care they seek in the hope of avoiding preterm births? Explain.

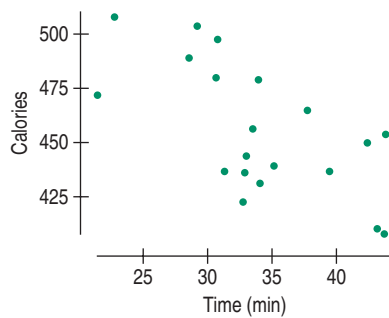


Preterm Birth Rate per 100 live twin births among U.S. twins by intensive, adequate, and less than adequate prenatal care utilization, 1981–1997. (*JAMA* 284[2000]: 335–341)



- T 33. Lunchtime.** Create and interpret a model for the toddlers' lunchtime data presented in Chapter 7. The table and graph show the number of minutes the kids stayed at the table and the number of calories they consumed.

Calories	Time	Calories	Time
472	21.4	450	42.4
498	30.8	410	43.1
465	37.7	504	29.2
456	33.5	437	31.3
423	32.8	489	28.6
437	39.5	436	32.9
508	22.8	480	30.6
431	34.1	439	35.1
479	33.9	444	33.0
454	43.8	408	43.7



- 34. Gasoline.** Since clean-air regulations have dictated the use of unleaded gasoline, the supply of leaded gas in New York state has diminished. The table below was given on the August 2001 New York State Math B exam, a statewide achievement test for high school students.

Year	1984	1988	1992	1996	2000
Gallons (1000's)	150	124	104	76	50

- Create a linear model to predict the number of gallons that will be available in 2005.
  - The exam then asked students to estimate the year when leaded gasoline will first become unavailable, expecting them to use the model from part a to answer the question. Explain why that method is incorrect.
  - Create a model that *would* be appropriate for that task, and make the estimate.
  - The "wrong" answer from the other model is fairly accurate in this case. *Why?*
- T 35. Tobacco and alcohol.** Are people who use tobacco products more likely to consume alcohol? Here are data on household spending (in pounds) taken by the British Government on 11 regions in Great Britain. Do tobacco and alcohol spending appear to be related? What questions do you have about these data? What conclusions can you draw?

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.20
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

- T 36. Football weights.** The Sears Cup was established in 1993 to honor institutions that maintain a broad-based athletic program, achieving success in many sports, both men's and women's. Since its Division III inception in 1995, the cup has been won by Williams College in every year except one. Their football team has a 85.3% winning record under their current coach. Why does the football team win so much? Is it because they're heavier than their opponents? The table shows the average team weights for selected years from 1973 to 1993.

Year	Weight (lb)	Year	Weight (lb)
1973	185.5	1983	192.0
1975	182.4	1987	196.9
1977	182.1	1989	202.9
1979	191.1	1991	206.0
1981	189.4	1993	198.7

- Fit a straight line to the relationship between *Weight* and *Year*.
  - Does a straight line seem reasonable?
  - Predict the average weight of the team for the year 2003. Does this seem reasonable?
  - What about the prediction for the year 2103? Explain.
  - What about the prediction for the year 3003? Explain.
- 37. Models.** Find the predicted value of  $y$ , using each model for  $x = 10$ .
- $\hat{y} = 2 + 0.8 \ln x$
  - $\log \hat{y} = 5 - 0.23x$
  - $\frac{1}{\sqrt{\hat{y}}} = 17.1 - 1.66x$
- T 38. Williams vs. Texas.** Here are the average weights of the football team for the University of Texas for various years in the 20th century.

Year	1905	1919	1932	1945	1955	1965
Weight (lb)	164	163	181	192	195	199

- Fit a straight line to the relationship of *Weight* by *Year* for Texas football players.
- According to these models, in what year will the predicted weight of the Williams College team from

Exercise 36 first be more than the weight of the University of Texas team?

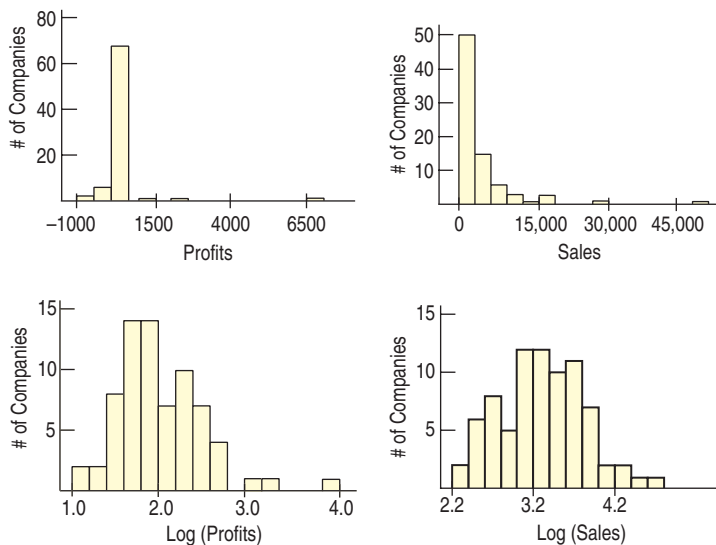
c) Do you believe this? Explain.

39. **Vehicle weights.** The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed “weigh-in-motion” scale. After installation of the scale, a study was conducted to find out whether the scale’s readings correspond to the true weights of the trucks being monitored. In Exercise 40 of Chapter 7, you examined the scatterplot for the data they collected, finding the association to be approximately linear with  $R^2 = 93\%$ . Their regression equation is  $\widehat{Wt} = 10.85 + 0.64 \text{ Scale}$ , where both the scale reading and the predicted weight of the truck are measured in thousands of pounds.

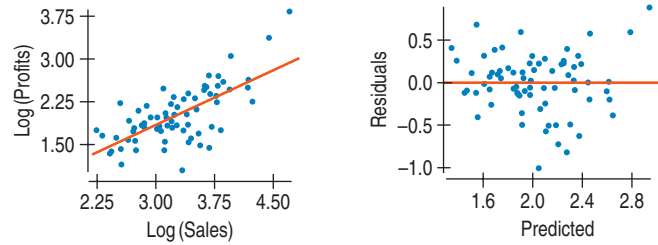
- a) Estimate the weight of a truck if this scale read 31,200 pounds.
- b) If that truck actually weighed 32,120 pounds, what was the residual?
- c) If the scale reads 35,590 pounds, and the truck has a residual of  $-2440$  pounds, how much does it actually weigh?
- d) In general, do you expect estimates made using this equation to be reasonably accurate? Explain.
- e) If the police plan to use this scale to issue tickets to trucks that appear to be overloaded, will negative or positive residuals be a greater problem? Explain.

40. **Profit.** How are a company’s profits related to its sales? Let’s examine data from 71 large U.S. corporations. All amounts are in millions of dollars.

- a) Histograms of *Profits* and *Sales* and histograms of the logarithms of *Profits* and *Sales* are on the next page. Why are the re-expressed data better for regression?



- b) Here are the scatterplot and residuals plot for the regression of logarithm of *Profits* vs. log of *Sales*. Do you think this model is appropriate? Explain.



c) Here’s the regression analysis. Write the equation.

Dependent variable is: Log Profit  
R-squared = 48.1%

Variable	Coefficient
Intercept	-0.106259
LogSales	0.647798

d) Use your equation to estimate profits earned by a company with sales of 2.5 billion dollars. (That’s 2500 million.)

T 41. **Down the drain.** Most water tanks have a drain plug so that the tank may be emptied when it’s to be moved or repaired. How long it takes a certain size of tank to drain depends on the size of the plug, as shown in the table. Create a model.

Plug Dia (in.)	$\frac{3}{8}$	$\frac{1}{2}$	$\frac{3}{4}$	1	$1\frac{1}{4}$	$1\frac{1}{2}$	2
Drain Time (min.)	140	80	35	20	13	10	5

T 42. **Chips.** A start-up company has developed an improved electronic chip for use in laboratory equipment. The company needs to project the manufacturing cost, so it develops a spreadsheet model that takes into account the purchase of production equipment, overhead, raw materials, depreciation, maintenance, and other business costs. The spreadsheet estimates the cost of producing 10,000 to 200,000 chips per year, as seen in the table. Develop a regression model to predict *Costs* based on the *Level* of production.

Chips Produced (1000s)	Cost per Chip (\$)	Chips Produced (1000s)	Cost per Chip (\$)
10	146.10	90	47.22
20	105.80	100	44.31
30	85.75	120	42.88
40	77.02	140	39.05
50	66.10	160	37.47
60	63.92	180	35.09
70	58.80	200	34.04
80	50.91		



PART



# Gathering Data

## Chapter 11

Understanding Randomness

## Chapter 12

Sample Surveys

## Chapter 13

Experiments and Observational Studies



# Understanding Randomness



*“The most decisive conceptual event of twentieth century physics has been the discovery that the world is not deterministic. . . . A space was cleared for chance.”*

— Ian Hacking,  
*The Taming of Chance*

**W**e all know what it means for something to be random. Or do we? Many children’s games rely on chance outcomes. Rolling dice, spinning spinners, and shuffling cards all select at random. Adult games use randomness as well, from card games to lotteries to Bingo. What’s the most important aspect of the randomness in these games? It must be fair.

What is it about random selection that makes it seem fair? It’s really two things. First, nobody can guess the outcome before it happens. Second, when we want things to be fair, usually some underlying set of outcomes will be equally likely (although in many games, some combinations of outcomes are more likely than others).

Randomness is not always what we might think of as “at random.” Random outcomes have a lot of structure, especially when viewed in the long run. You can’t predict how a fair coin will land on any single toss, but you’re pretty confident that if you flipped it thousands of times you’d see about 50% heads. As we will see, randomness is an essential tool of Statistics. Statisticians don’t think of randomness as the annoying tendency of things to be unpredictable or haphazard. Statisticians use randomness as a tool. In fact, without deliberately applying randomness, we couldn’t do most of Statistics, and this book would stop right about here.<sup>1</sup>

But truly random values are surprisingly hard to get. Just to see how fair humans are at selecting, pick a number at random from the top of the next page. Go ahead. Turn the page, look at the numbers quickly, and pick a number at random.

Ready?

Go.

<sup>1</sup> Don’t get your hopes up.

# 1 2 3 4

## It's Not Easy Being Random

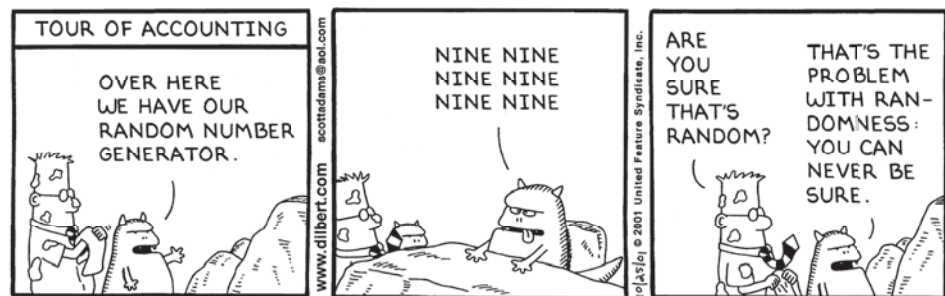
*"The generation of random numbers is too important to be left to chance."*

—Robert R. Coveyou,  
Oak Ridge National  
Laboratory

**A S** **Activity: Random Behavior.** *ActivStats*' Random Experiment Tool lets you experiment with truly random outcomes. We'll use it a lot in the coming chapters.

Did you pick 3? If so, you've got company. Almost 75% of all people pick the number 3. About 20% pick either 2 or 4. If you picked 1, well, consider yourself a little different. Only about 5% choose 1. Psychologists have proposed reasons for this phenomenon, but for us, it simply serves as a lesson that we've got to find a better way to choose things at random.

So how should we generate **random numbers**? It's surprisingly difficult to get random values even when they're equally likely. Computers have become a popular way to generate random numbers. Even though they often do much better than humans, computers can't generate truly random numbers either. Computers follow programs. Start a computer from the same place, and it will always follow exactly the same path. So numbers generated by a computer program are not truly random. Technically, "random" numbers generated this way are *pseudorandom* numbers. Pseudorandom values are generated in a fixed sequence, and because computers can represent only a finite number of distinct values, the sequence of pseudorandom numbers must eventually repeat itself. Fortunately, pseudorandom values are good enough for most purposes because they are virtually indistinguishable from truly random numbers.



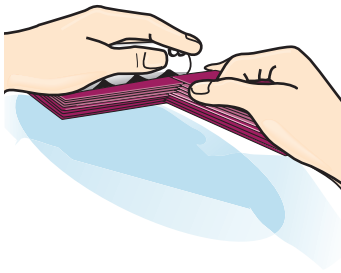
**A S** **Activity: Truly Random Values on the Internet.** This activity will take you to an Internet site ([www.random.org](http://www.random.org)) that generates all the truly random numbers you could want.

There *are* ways to generate random numbers so that they are both equally likely and truly random. In the past, entire books of carefully generated random numbers were published. The books never made the best-seller lists and probably didn't make for great reading, but they were quite valuable to those who needed truly random values.<sup>2</sup> Today, we have a choice. We can use these books or find genuinely random digits from several Internet sites. The sites use methods like timing the decay of a radioactive element or even the random changes of lava

<sup>2</sup> You'll find a table of random digits of this kind in the back of this book.



An ordinary deck of playing cards, like the ones used in bridge and many other card games, consists of 52 cards. There are numbered cards (2 through 10), and face cards (Jack, Queen, King, Ace) whose value depends on the game you are playing. Each card is also marked by one of four suits (clubs, diamonds, hearts, or spades) whose significance is also game-specific.



lamps to generate truly random digits.<sup>3</sup> In either case, a string of random digits might look like this:

```
2217726304387410092537086270581997622725849795907032825001108963
3217535822643800292254644943760642389043766557204107354186024508
8906427308645681412198226653885873285801699027843110380420067664
8740522639824530519902027044464984322000946238678577902639002954
8887003319933147508331265192321413908608674496383528968974910533
6944182713168919406022181281304751019321546303870481407676636740
6070204916508913632855351361361043794293428486909462881431793360
7706356513310563210508993624272872250535395513645991015328128202
```

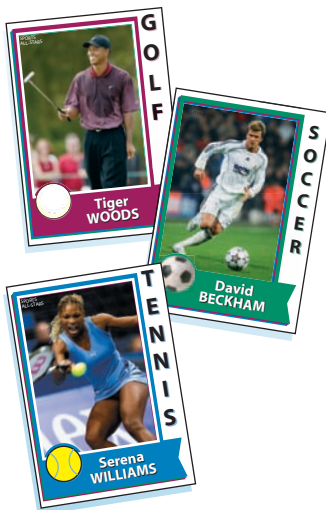
You probably have more interesting things to download than a few million random digits, but we'll discuss ways to use such random digits to apply randomness to real situations soon. The best ways we know to generate data that give a fair and accurate picture of the world rely on randomness, and the ways in which we draw conclusions from those data depend on the randomness, too.

**Aren't you done shuffling yet?** Even something as common as card shuffling may not be as random as you might think. If you shuffle cards by the usual method in which you split the deck in half and try to let cards fall roughly alternately from each half, you're doing a "riffle shuffle."

How many times should you shuffle cards to make the deck random? A surprising fact was discovered by statisticians Persi Diaconis, Ronald Graham, and W. M. Kantor. It takes seven riffle shuffles. Fewer than seven leaves order in the deck, but after that, more shuffling does little good. Most people, though, don't shuffle that many times.

When computers were first used to generate hands in bridge tournaments, some professional bridge players complained that the computer was making too many "weird" hands—hands with 10 cards of one suit, for example. Suddenly these hands were appearing more often than players were used to when cards were shuffled by hand. The players assumed that the computer was doing something wrong. But it turns out that it's humans who hadn't been shuffling enough to make the decks really random and have those "weird" hands appear as often as they should.

## Practical Randomness



Suppose a cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal in the hope of boosting sales. The manufacturer announces that 20% of the boxes contain a picture of Tiger Woods, 30% a picture of David Beckham, and the rest a picture of Serena Williams. You want all three pictures. How many boxes of cereal do you expect to have to buy in order to get the complete set?

How can we answer questions like this? Well, one way is to buy hundreds of boxes of cereal to see what might happen. But let's not. Instead, we'll consider using a random model. Why random? When we pick a box of cereal off the shelf, we don't know what picture is inside. We'll assume that the pictures are randomly placed in the boxes and that the boxes are distributed randomly to stores around the country. Why a model? Because we won't actually buy the cereal boxes. We can't afford all those boxes and we don't want to waste food. So we need an imitation of the real process that we can manipulate and control. In short, we're going to **simulate** reality.

<sup>3</sup> For example, [www.random.org](http://www.random.org) or [www.randomnumbers.info](http://www.randomnumbers.info).

## A Simulation

Modern physics has shown that randomness is not just a mathematical game; it is fundamentally the way the universe works.

*Regardless of improvements in data collection or in computer power, the best we can ever do, according to quantum mechanics . . . is predict the probability that an electron, or a proton, or a neutron, or any other of nature's constituents, will be found here or there. Probability reigns supreme in the microcosmos.*

—Brian Greene, *The Fabric of the Cosmos: Space, Time, and the Texture of Reality* (p. 91)

The question we've asked is how many boxes do you expect to buy to get a complete card collection. But we can't answer our question by completing a card collection just once. We want to understand the *typical* number of boxes to open, how that number varies, and, often, the shape of the distribution. So we'll have to do this over and over. We call each time we obtain a simulated answer to our question a **trial**.

For the sports cards, a trial's outcome is the number of boxes. We'll need at least 3 boxes to get one of each card, but with really bad luck, you could empty the shelves of several supermarkets before finding the card you need to get all 3. So, the possible outcomes of a trial are 3, 4, 5, or lots more. But we can't simply pick one of those numbers at random, because they're not equally likely. We'd be surprised if we only needed 3 boxes to get all the cards, but we'd probably be even more surprised to find that it took exactly 7,359 boxes. In fact, the reason we're doing the simulation is that it's hard to guess how many boxes we'd expect to open.

### BUILDING A SIMULATION

We know how to find equally likely random digits. How can we get from there to simulating the trial outcomes? We know the relative frequencies of the cards: 20% Tiger, 30% Beckham, and 50% Serena. So, we can interpret the digits 0 and 1 as finding Tiger; 2, 3, and 4 as finding Beckham; and 5 through 9 as finding Serena to simulate opening one box. Opening one box is the basic building block, called a **component** of our simulation. But the component's outcome isn't the result we want. We need to observe a sequence of components until our card collection is complete. The *trial's* outcome is called the **response variable**; for this simulation that's the *number* of components (boxes) in the sequence.

Let's look at the steps for making a simulation:

#### Specify how to model a component outcome using equally likely random digits:

1. **Identify the component to be repeated.** In this case, our component is the opening of a box of cereal.
2. **Explain how you will model the component's outcome.** The digits from 0 to 9 are equally likely to occur. Because 20% of the boxes contain Tiger's picture, we'll use 2 of the 10 digits to represent that outcome. Three of the 10 digits can model the 30% of boxes with David Beckham cards, and the remaining 5 digits can represent the 50% of boxes with Serena. One possible assignment of the digits, then, is

0, 1 Tiger   2, 3, 4 Beckham   5, 6, 7, 8, 9 Serena.

#### Specify how to simulate trials:

3. **Explain how you will combine the components to model a trial.** We pretend to open boxes (repeat components) until our collection is complete. We do this by looking at each random digit and indicating what picture it represents. We continue until we've found all three.
4. **State clearly what the response variable is.** What are we interested in? We want to find out the number of boxes it might take to get all three pictures.

#### Put it all together to run the simulation:

5. **Run several trials.** For example, consider the third line of random digits shown earlier (p. 257):

8906427308645681412198226653885873285801699027843110380420067664.

Let's see what happened.



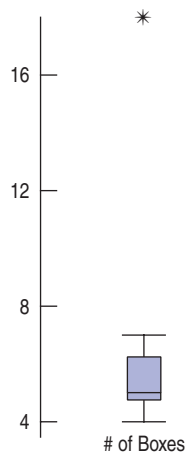
The first random digit, 8, means you get Serena's picture. So the first component's outcome is Serena. The second digit, 9, means Serena's picture is also in the next box. Continuing to interpret the random digits, we get Tiger's picture (0) in the third, Serena's (6) again in the fourth, and finally Beckham (4) on the fifth box. Since we've now found all three pictures, we've finished one trial of our simulation. This trial's outcome is 5 boxes.

Now we keep going, running more trials by looking at the rest of our line of random digits:

89064 2730 8645681 41219 822665388587328580 169902 78431 1038 042006 7664.

It's best to create a chart to keep track of what happens:

Trial Number	Component Outcomes	Trial Outcomes: $y$ = Number of boxes
1	89064 = <b>Serena</b> , Serena, <b>Tiger</b> , Serena, <b>Beckham</b>	5
2	2730 = <b>Beckham</b> , <b>Serena</b> , Beckham, <b>Tiger</b>	4
3	8645681 = <b>Serena</b> , Serena, <b>Beckham</b> , . . . , <b>Tiger</b>	7
4	41219 = <b>Beckham</b> , <b>Tiger</b> , Beckham, Tiger, <b>Serena</b>	5
5	822665388587328580 = <b>Serena</b> , <b>Beckham</b> , . . . , <b>Tiger</b>	18
6	169902 = <b>Tiger</b> , <b>Serena</b> , Serena, Serena, Tiger, <b>Beckham</b>	6
7	78431 = <b>Serena</b> , Serena, <b>Beckham</b> , Beckham, <b>Tiger</b>	5
8	1038 = <b>Tiger</b> , Tiger, <b>Beckham</b> , <b>Serena</b>	4
9	042006 = <b>Tiger</b> , <b>Beckham</b> , Beckham, Tiger, Tiger, <b>Serena</b>	6
10	7664 . . . = <b>Serena</b> , Serena, Serena, <b>Beckham</b> . . .	?



**A S**

**Activity: Bigger Samples Are Better.** The random simulation tool can generate lots of outcomes with a single click, so you can see more of the long run with less effort.

#### Analyze the response variable:

- Collect and summarize the results of all the trials.** You know how to summarize and display a response variable. You'll certainly want to report the shape, center, and spread, and depending on the question asked, you may want to include more.
- State your conclusion,** as always, in the context of the question you wanted to answer. Based on this simulation, we estimate that customers hoping to complete their card collection will need to open a median of 5 boxes, but it could take a lot more.

If you fear that these may not be accurate estimates because we ran only nine trials, you are absolutely correct. The more trials the better, and nine is woefully inadequate. Twenty trials is probably a reasonable minimum if you are doing this by hand. Even better, use a computer and run a few hundred trials.

## FOR EXAMPLE

### Simulating a dice game

The game of 21 can be played with an ordinary 6-sided die. Competitors each roll the die repeatedly, trying to get the highest total less than or equal to 21. If your total exceeds 21, you lose.

Suppose your opponent has rolled an 18. Your task is to try to beat him by getting more than 18 points without going over 21. How many rolls do you expect to make, and what are your chances of winning?

**Question:** How will you simulate the components?

A component is one roll of the die. I'll simulate each roll by looking at a random digit from a table or an Internet site. The digits 1 through 6 will represent the results on the die; I'll ignore digits 7–9 and 0.

(continued)

For Example (continued)

**Question:** How will you combine components to model a trial? What's the response variable?

I'll add components until my total is greater than 18, counting the number of rolls. If my total is greater than 21, it is a loss; if not, it is a win. There are two response variables. I'll count the number of times I roll the die, and I'll keep track of whether I win or lose.

**Question:** How would you use these random digits to run trials? Show your method clearly for two trials.

91129 58757 69274 92380 82464 33089

I've marked the discarded digits in color.

Trial #1:	9	1	1	2	9	5	8	7	5	7	6			
Total:		1	2	4		9			14		20	Outcomes: 6 rolls, won		
Trial #2:	9	2	7	4	9	2	3	8	0	8	2	4	6	
Total:		2		6		8	11			13	17	23	Outcomes: 7 rolls, lost	

**Question:** Suppose you run 30 trials, getting the outcomes tallied here. What is your conclusion?

Based on my simulation, when competing against an opponent who has a score of 18, I expect my turn to usually last 5 or 6 rolls, and I should win about 70% of the time.

Number of rolls	Result
4	Won IIII
5	Lost IIII
6	IIII
7	IIII
8	I



### JUST CHECKING

The baseball World Series consists of up to seven games. The first team to win four games wins the series. The first two are played at one team's home ballpark, the next three at the other team's park, and the final two (if needed) are played back at the first park. Records over the past century show that there is a home field advantage; the home team has about a 55% chance of winning. Does the current system of alternating ballparks even out the home field advantage? How often will the team that begins at home win the series?

Let's set up the simulation:

1. What is the component to be repeated?
2. How will you model each component from equally likely random digits?
3. How will you model a trial by combining components?
4. What is the response variable?
5. How will you analyze the response variable?

### STEP-BY-STEP EXAMPLE

#### Simulation

Fifty-seven students participated in a lottery for a particularly desirable dorm room—a triple with a fireplace and private bath in the tower. Twenty of the participants were members of the same varsity team. When all three winners were members of the team, the other students cried foul.

**Question:** Could an all-team outcome reasonably be expected to happen if everyone had a fair shot at the room?



**Plan** State the problem. Identify the important parts of your simulation.

**Components** Identify the components.

**Outcomes** State how you will model each component using equally likely random digits. You can't just use the digits from 0 to 9 because the outcomes you are simulating are not multiples of 10%.

There are 20 and 37 students in the two groups. This time you must use *pairs* of random digits (and ignore some of them) to represent the 57 students.

**Trial** Explain how you will combine the components to simulate a trial. In each of these trials, you can't choose the same student twice, so you'll need to ignore a random number if it comes up a second or third time. Be sure to mention this in describing your simulation.

**Response Variable** Define your response variable.

I'll use a simulation to investigate whether it's unlikely that three varsity athletes would get the great room in the dorm if the lottery were fair.

A component is the selection of a student.

I'll look at two-digit random numbers.

Let 00–19 represent the 20 varsity applicants.

Let 20–56 represent the other 37 applicants.

Skip 57–99. If I get a number in this range, I'll throw it away and go back for another two-digit random number.

Each trial consists of identifying pairs of digits as V (varsity) or N (nonvarsity) until 3 people are chosen, ignoring out-of-range or repeated numbers (X)—I can't put the same person in the room twice.

The response variable is whether or not all three selected students are on the varsity team.



**Mechanics** Run several trials. Carefully record the random numbers, indicating

- 1) the corresponding component outcomes (here, Varsity, Nonvarsity, or ignored number) and
- 2) the value of the response variable.

Trial Number	Component Outcomes	All Varsity?
1	74 02 94 39 02 77 55 X V X N X X N	No
2	18 63 33 25 V X N N	No
3	05 45 88 91 56 V N X X N	No
4	39 09 07 N V V	No
5	65 39 45 95 43 X N N X N	No
6	98 95 11 68 77 12 17 X X V X X V V	Yes
7	26 19 89 93 77 27 N V X X X N	No

(continued)

**Analyze** Summarize the results across all trials to answer the initial question.

**TELL** **Conclusion** Describe what the simulation shows, and interpret your results in the context of the real world.

8	23 52 37 N N N	No
9	16 50 83 44 V N X N	No
10	74 17 46 85 09 X V N X V	No

“All varsity” occurred once, or 10% of the time.

In my simulation of “fair” room draws, the three people chosen were all varsity team members only 10% of the time. While this result could happen by chance, it is not particularly likely. I’m suspicious, but I’d need many more trials and a smaller frequency of the all-varsity outcome before I would make an accusation of unfairness.

**TI Tips**

**Generating random numbers**

```
MATH NUM CPX PRB
1:rand
2:nPr
3:nCr
4:!
5:randInt(
6:randNorm(
7:randBin(
```

```
randInt(0,1) 0
randInt(1,6) 2
```

```
randInt(1,6,2)
(2 1)
(3 2)
(6 4)
(2 5)
(2 6)
(5 1)
```

```
randInt(0,9,5)
(0 6 0 5 9)
```

```
randInt(0,56,3)
(14 14 35)
(50 17 45)
(36 25 10)
(33 24 19)
(0 12 26)
(33 11 19)
```

Instead of using coins, dice, cards, or tables of random numbers, you may decide to use your calculator for simulations. There are several random number generators offered in the **MATH PRB** menu.

**5:randInt()** is of particular importance. This command will produce any number of random integers in a specified range.

Here are some examples showing how to use **randInt** for simulations:

- **randInt(0,1)** randomly chooses a 0 or a 1. This is an effective simulation of a coin toss. You could let 0 represent tails and 1 represent heads.
- **randInt(1,6)** produces a random integer from 1 to 6, a good way to simulate rolling a die.
- **randInt(1,6,2)** simulates rolling *two* dice. To do several rolls in a row, just hit **ENTER** repeatedly.
- **randInt(0,9,5)** produces five random integers that might represent the pictures in the cereal boxes. Our run gave us two Tigers (0, 1), no Beckhams (2, 3, 4), and three Serenas (5–9).
- **randInt(0,56,3)** produces three random integers between 0 and 56, a nice way to simulate the dorm room lottery. The window shows 6 trials, but we would skip the first one because one student was chosen twice. In none of the remaining 5 trials did three athletes (0–19) win.

## WHAT CAN GO WRONG?

**A S** **Activity: Estimating Summaries from Random Outcomes.** See how well you can estimate something you can't know just by generating random outcomes.

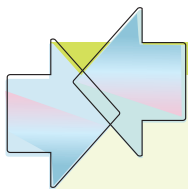
**TI-nspire**  
**Simulations.** Improve your predictions by running thousands of trials.

- ▶ **Don't overstate your case.** Let's face it: In some sense, a simulation is *always* wrong. After all, it's not the real thing. We didn't buy any cereal or run a room draw. So beware of confusing what *really* happens with what a simulation suggests *might* happen. Never forget that future results will not match your simulated results exactly.
- ▶ **Model outcome chances accurately.** A common mistake in constructing a simulation is to adopt a strategy that may appear to produce the right kind of results, but that does not accurately model the situation. For example, in our room draw, we could have gotten 0, 1, 2, or 3 team members. Why not just see how often these digits occur in random digits from 0 to 9, ignoring the digits 4 and up?

3 2 1 7 9 0 0 5 9 7 3 7 9 2 5 2 4 1 3 8  
 3 2 1 x x 0 0 x x x 3 x x 2 x 2 x 1 3 x

This "simulation" makes it seem fairly likely that three team members would be chosen. There's a big problem with this approach, though: The digits 0, 1, 2, and 3 occur with equal frequency among random digits, making each outcome appear to happen 25% of the time. In fact, the selection of 0, 1, 2, or all 3 team members are not all equally likely outcomes. In our correct simulation, we estimated that all 3 would be chosen only about 10% of the time. If your simulation overlooks important aspects of the real situation, your model will not be accurate.

- ▶ **Run enough trials.** Simulation is cheap and fairly easy to do. Don't try to draw conclusions based on 5 or 10 trials (even though we did for illustration purposes here). We'll make precise how many trials to use in later chapters. For now, err on the side of large numbers of trials.



## CONNECTIONS

Simulations often generate many outcomes of a response variable, and we are often interested in the distribution of these responses. The tools we use to display and summarize the distribution of any real variable are appropriate for displaying and summarizing randomly generated responses as well.

Make histograms, boxplots, and Normal probability plots of the response variables from simulations, and summarize them with measures of center and spread. Be especially careful to report the variation of your response variable.

Don't forget to think about your analyses. Simulations can hide subtle errors. A careful analysis of the responses can save you from erroneous conclusions based on a faulty simulation.

You may be less likely to find an outlier in simulated responses, but if you find one, you should certainly determine how it happened.

## WHAT HAVE WE LEARNED?



We've learned to harness the power of randomness. We've learned that a simulation model can help us investigate a question for which many outcomes are possible, we can't (or don't want to) collect data, and a mathematical answer is hard to calculate. We've learned how to base our simulation on random values generated by a computer, generated by a randomizing device such as a die or spinner, or found on the Internet. Like all models, simulations can provide us with useful insights about the real world.

## Terms

Random	255. An outcome is random if we know the possible values it can have, but not which particular value it takes.
Generating random numbers	256. Random numbers are hard to generate. Nevertheless, several Internet sites offer an unlimited supply of equally likely random values.
Simulation	258. A simulation models a real-world situation by using random-digit outcomes to mimic the uncertainty of a response variable of interest.
Simulation component	258. A component uses equally likely random digits to model simple random occurrences whose outcomes may not be equally likely.
Trial	258. The sequence of several components representing events that we are pretending will take place.
Response variable	258. Values of the response variable record the results of each trial with respect to what we were interested in.

## Skills



- ▶ Be able to recognize random outcomes in a real-world situation.
- ▶ Be able to recognize when a simulation might usefully model random behavior in the real world.



- ▶ Know how to perform a simulation either by generating random numbers on a computer or calculator, or by using some other source of random values, such as dice, a spinner, or a table of random numbers.



- ▶ Be able to describe a simulation so that others can repeat it.
- ▶ Be able to discuss the results of a simulation study and draw conclusions about the question being investigated.

## SIMULATION ON THE COMPUTER

Simulations are best done with the help of technology simply because more trials makes a better simulation, and computers are fast. There are special computer programs designed for simulation, and most statistics packages and calculators can at least generate random numbers to support a simulation.

All technology-generated random numbers are *pseudorandom*. The random numbers available on the Internet may technically be better, but the differences won't matter for any simulation of modest size. Pseudorandom numbers generate the next random value from the previous one by a specified algorithm. But they have to start somewhere. This starting point is called the "seed." Most programs let you set the seed. There's usually little reason to do this, but if you wish to, go ahead. If you reset the seed to the same value, the programs will generate the same sequence of "random" numbers.

**AS** **Activity: Creating Random Values.** Learn to use your statistics package to generate random outcomes.



## EXERCISES

- Coin toss.** Is a coin flip random? Why or why not?
- Casino.** A casino claims that its electronic “video roulette” machine is truly random. What should that claim mean?
- The lottery.** Many states run lotteries, giving away millions of dollars if you match a certain set of winning numbers. How are those numbers determined? Do you think this method guarantees randomness? Explain.
- Games.** Many kinds of games people play rely on randomness. Cite three different methods commonly used in the attempt to achieve this randomness, and discuss the effectiveness of each.
- Birth defects.** The American College of Obstetricians and Gynecologists says that out of every 100 babies born in the United States, 3 have some kind of major birth defect. How would you assign random numbers to conduct a simulation based on this statistic?
- Colorblind.** By some estimates, about 10% of all males have some color perception defect, most commonly red-green colorblindness. How would you assign random numbers to conduct a simulation based on this statistic?
- Geography.** An elementary school teacher with 25 students plans to have each of them make a poster about two different states. The teacher first numbers the states (in alphabetical order, from 1-Alabama to 50-Wyoming), then uses a random number table to decide which states each kid gets. Here are the random digits:  
45921 01710 22892 37076
  - Which two state numbers does the first student get?
  - Which two state numbers go to the second student?
- Get rich.** Your state’s BigBucks Lottery prize has reached \$100,000,000, and you decide to play. You have to pick five numbers between 1 and 60, and you’ll win if your numbers match those drawn by the state. You decide to pick your “lucky” numbers using a random number table. Which numbers do you play, based on these random digits?  
43680 98750 13092 76561 58712
  - Describe how you will simulate a component.
  - Describe how you will simulate a trial.
  - Describe the response variable.
- Play the lottery.** Some people play state-run lotteries by always playing the same favorite “lucky” number. Assuming that the lottery is truly random, is this strategy better, worse, or the same as choosing different numbers for each play? Explain.
- Play it again, Sam.** In Exercise 8 you imagined playing the lottery by using random digits to decide what numbers to play. Is this a particularly good or bad strategy? Explain.
- Bad simulations.** Explain why each of the following simulations fails to model the real situation properly:
  - Use a random integer from 0 through 9 to represent the number of heads when 9 coins are tossed.
  - A basketball player takes a foul shot. Look at a random digit, using an odd digit to represent a good shot and an even digit to represent a miss.
  - Use random digits from 1 through 13 to represent the denominations of the cards in a five-card poker hand.
- More bad simulations.** Explain why each of the following simulations fails to model the real situation:
  - Use random numbers 2 through 12 to represent the sum of the faces when two dice are rolled.
  - Use a random integer from 0 through 5 to represent the number of boys in a family of 5 children.
  - Simulate a baseball player’s performance at bat by letting 0 = an out, 1 = a single, 2 = a double, 3 = a triple, and 4 = a home run.
- Wrong conclusion.** A Statistics student properly simulated the length of checkout lines in a grocery store and then reported, “The average length of the line will be 3.2 people.” What’s wrong with this conclusion?
- Another wrong conclusion.** After simulating the spread of a disease, a researcher wrote, “24% of the people contracted the disease.” What should the correct conclusion be?
- Election.** You’re pretty sure that your candidate for class president has about 55% of the votes in the entire school. But you’re worried that only 100 students will show up to vote. How often will the underdog (the one with 45% support) win? To find out, you set up a simulation.
  - Describe how you will simulate a component.
  - Describe how you will simulate a trial.
  - Describe the response variable.
- Two pair or three of a kind?** When drawing five cards randomly from a deck, which is more likely, two pairs or three of a kind? A pair is exactly two of the same denomination. Three of a kind is exactly 3 of the same denomination. (Don’t count three 8’s as a pair—that’s 3 of a kind. And don’t count 4 of the same kind as two pair—that’s 4 of a kind, a very special hand.) How could you simulate 5-card hands? Be careful; once you’ve picked the 8 of spades, you can’t get it again in that hand.
  - Describe how you will simulate a component.
  - Describe how you will simulate a trial.
  - Describe the response variable.
- Cereal.** In the chapter’s example, 20% of the cereal boxes contained a picture of Tiger Woods, 30% David Beckham, and the rest Serena Williams. Suppose you buy five boxes of cereal. Estimate the probability that you end up with a complete set of the pictures. Your simulation should have at least 20 runs.
- Cereal, again.** Suppose you really want the Tiger Woods picture. How many boxes of cereal do you need to buy to be pretty sure of getting at least one? Your simulation should use at least 10 trials.

19. **Multiple choice.** You take a quiz with 6 multiple choice questions. After you studied, you estimated that you would have about an 80% chance of getting any individual question right. What are your chances of getting them all right? Use at least 20 trials.
20. **Lucky guessing?** A friend of yours who took the multiple choice quiz in Exercise 19 got all 6 questions right, but now claims to have guessed blindly on every question. If each question offered 4 possible answers, do you believe her? Explain, basing your argument on a simulation involving at least 10 trials.
21. **Beat the lottery.** Many states run lotteries to raise money. A Web site advertises that it knows “how to increase YOUR chances of Winning the Lottery.” They offer several systems and criticize others as foolish. One system is called *Lucky Numbers*. People who play the *Lucky Numbers* system just pick a “lucky” number to play, but maybe some numbers are luckier than others. Let’s use a simulation to see how well this system works.  
To make the situation manageable, simulate a simple lottery in which a single digit from 0 to 9 is selected as the winning number. Pick a single value to bet, such as 1, and keep playing it over and over. You’ll want to run at least 100 trials. (If you can program the simulations on a computer, run several hundred. Or generalize the questions to a lottery that chooses two- or three-digit numbers—for which you’ll need thousands of trials.)  
a) What proportion of the time do you expect to win?  
b) Would you expect better results if you picked a “luckier” number, such as 7? (Try it if you don’t know.) Explain.
22. **Random is as random does.** The “beat the lottery” Web site discussed in Exercise 21 suggests that because lottery numbers are random, it is better to select your bet randomly. For the same simple lottery in Exercise 21 (random values from 0 to 9), generate each bet by choosing a separate random value between 0 and 9. Play many games. What proportion of the time do you win?
23. **It evens out in the end.** The “beat the lottery” Web site of Exercise 21 notes that in the long run we expect each value to turn up about the same number of times. That leads to their recommended strategy. First, watch the lottery for a while, recording the winners. Then bet the value that has turned up the least, because it will need to turn up more often to even things out. If there is more than one “rarest” value, just take the lowest one (since it doesn’t matter). Simulating the simplified lottery described in Exercise 21, play many games with this system. What proportion of the time do you win?
24. **Play the winner?** Another strategy for beating the lottery is the reverse of the system described in Exercise 23. Simulate the simplified lottery described in Exercise 21. Each time, bet the number that just turned up. The Web site suggests that this method should do worse. Does it? Play many games and see.
25. **Driving test.** You are about to take the road test for your driver’s license. You hear that only 34% of candidates pass the test the first time, but the percentage rises to 72% on subsequent retests. Estimate the average number of tests drivers take in order to get a license. Your simulation should use at least 20 runs.
26. **Still learning?** As in Exercise 25, assume that your chance of passing the driver’s test is 34% the first time and 72% for subsequent retests. Estimate the percentage of those tested who still do not have a driver’s license after two attempts.
27. **Basketball strategy.** Late in a basketball game, the team that is behind often fouls someone in an attempt to get the ball back. Usually the opposing player will get to shoot foul shots “one and one,” meaning he gets a shot, and then a second shot only if he makes the first one. Suppose the opposing player has made 72% of his foul shots this season. Estimate the number of points he will score in a one-and-one situation.
28. **Blood donors.** A person with type O-positive blood can receive blood only from other type O donors. About 44% of the U.S. population has type O blood. At a blood drive, how many potential donors do you expect to examine in order to get three units of type O blood?
29. **Free groceries.** To attract shoppers, a supermarket runs a weekly contest that involves “scratch-off” cards. With each purchase, customers get a card with a black spot obscuring a message. When the spot is scratched away, most of the cards simply say, “Sorry—please try again.” But during the week, 100 customers will get cards that make them eligible for a drawing for free groceries. Ten of the cards say they may be worth \$200, 10 others say \$100, 20 may be worth \$50, and the rest could be worth \$20. To register those cards, customers write their names on them and put them in a barrel at the front of the store. At the end of the week the store manager draws cards at random, awarding the lucky customers free groceries in the amount specified on their card. The drawings continue until the store has given away more than \$500 of free groceries. Estimate the average number of winners each week.
30. **Find the ace.** A new electronics store holds a contest to attract shoppers. Once an hour someone in the store is chosen at random to play the Music Game. Here’s how it works: An ace and four other cards are shuffled and placed face down on a table. The customer gets to turn cards over one at a time, looking for the ace. The person wins \$100 worth of free CDs or DVDs if the ace is the first card, \$50 if it is the second card, and \$20, \$10, or \$5 if it is the third, fourth, or fifth card chosen. What is the average dollar amount of music the store will give away?
31. **The family.** Many couples want to have both a boy and a girl. If they decide to continue to have children until they have one child of each sex, what would the average family size be? Assume that boys and girls are equally likely.
32. **A bigger family.** Suppose a couple will continue having children until they have at least two children of each sex (two boys *and* two girls). How many children might they expect to have?



33. **Dice game.** You are playing a children's game in which the number of spaces you get to move is determined by the rolling of a die. You must land exactly on the final space in order to win. If you are 10 spaces away, how many turns might it take you to win?
34. **Parcheesi.** You are three spaces from a win in Parcheesi. On each turn, you will roll two dice. To win, you must roll a total of 3 or roll a 3 on one of the dice. How many turns might you expect this to take?
35. **The hot hand.** A basketball player with a 65% shooting percentage has just made 6 shots in a row. The announcer says this player "is hot tonight! She's in the zone!" Assume the player takes about 20 shots per game. Is it unusual for her to make 6 or more shots in a row during a game?
36. **The World Series.** The World Series ends when a team wins 4 games. Suppose that sports analysts consider one team a bit stronger, with a 55% chance to win any individual game. Estimate the likelihood that the underdog wins the series.
37. **Teammates.** Four couples at a dinner party play a board game after the meal. They decide to play as teams of two and to select the teams randomly. All eight people write their names on slips of paper. The slips are thoroughly mixed, then drawn two at a time. How likely is it that every person will be teamed with someone other than the person he or she came to the party with?
38. **Second team.** Suppose the couples in Exercise 37 choose the teams by having one member of each couple write their names on the cards and the other people each pick a card at random. How likely is it that every person will be teamed with someone other than the person he or she came with?
39. **Job discrimination?** A company with a large sales staff announces openings for three positions as regional managers. Twenty-two of the current salespersons apply, 12 men and 10 women. After the interviews, when the company announces the newly appointed managers, all three positions go to women. The men complain of job discrimination. Do they have a case? Simulate a random selection of three people from the applicant pool, and make a decision about the likelihood that a fair process would result in hiring all women.
40. **Cell phones.** A proud legislator claims that your state's new law against talking on a cell phone while driving has reduced cell phone use to less than 12% of all drivers. While waiting for your bus the next morning, you notice that 4 of the 10 people who drive by are using their cell phones. Does this cast doubt on the legislator's figure of 12%? Use a simulation to estimate the likelihood of seeing at least 4 of 10 randomly selected drivers talking on their cell phones if the actual rate of usage is 12%. Explain your conclusion clearly.



### JUST CHECKING Answers

1. The component is one game.
2. I'll generate random numbers and assign numbers from 00 to 54 to the home team's winning and from 55 to 99 to the visitors' winning.
3. I'll generate components until one team wins 4 games. I'll record which team wins the series.
4. The response is who wins the series.
5. I'll calculate the proportion of wins by the team that starts at home.

# Sample Surveys



In 2007, Pew Research conducted a survey to assess Americans' knowledge of current events. They asked a random sample of 1,502 U.S. adults 23 factual questions about topics currently in the news.<sup>1</sup> Pew also asked respondents where they got their news. Those who frequented major newspaper Web sites or who are regular viewers of the *Daily Show* or *Colbert Report* scored best on knowledge of current events.<sup>2</sup> Even among those viewers, only 54% responded correctly to 15 or more of the questions. Pew claimed that this was close to the true percentage responding correctly that they would have found if they had asked all U.S. adults who got their news from those sources. That step from a small sample to the entire population is impossible without understanding Statistics. To make business decisions, to do science, to choose wise investments, or to understand what voters think they'll do the next election, we need to stretch beyond the data at hand to the world at large.

To make that stretch, we need three ideas. You'll find the first one natural. The second may be more surprising. The third is one of the strange but true facts that often confuse those who don't know Statistics.

## Idea 1: Examine a Part of the Whole

**A S** **Activity: Populations and Samples.** Explore the differences between populations and samples.

The first idea is to draw a sample. We'd like to know about an entire **population** of individuals, but examining all of them is usually impractical, if not impossible. So we settle for examining a smaller group of individuals—a **sample**—selected from the population.

You do this every day. For example, suppose you wonder how the vegetable soup you're cooking for dinner tonight is going to go over with your friends. To decide whether it meets your standards, you only need to try a small amount. You might taste just a spoonful or two. You certainly don't have to consume the whole

<sup>1</sup> For example, two of the questions were "Who is the vice-president of the United States?" and "What party controls Congress?"

<sup>2</sup> The lowest scores came from those whose main source of news was network morning shows or *Fox News*.

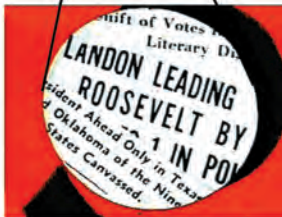
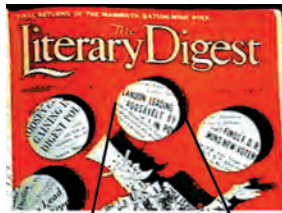
### The W's and Sampling

The population we are interested in is usually determined by the *Why* of our study. The sample we draw will be the *Who*. *When* and *How* we draw the sample may depend on what is practical.

pot. You trust that the taste will *represent* the flavor of the entire pot. The idea behind your tasting is that a small sample, if selected properly, can represent the entire population.

It's hard to go a day without hearing about the latest opinion poll. These polls are examples of **sample surveys**, designed to ask questions of a small group of people in the hope of learning something about the entire population. Most likely, you've never been selected to be part of one of these national opinion polls. That's true of most people. So how can the pollsters claim that a sample is representative of the entire population? The answer is that professional pollsters work quite hard to ensure that the "taste"—the sample that they take—represents the population. If not, the sample can give misleading information about the population.

## Bias



In 1936, a young pollster named George Gallup used a subsample of only 3000 of the 2.4 million responses that the *Literary Digest* received to reproduce the wrong prediction of Landon's victory over Roosevelt. He then used an entirely different sample of 50,000 and predicted that Roosevelt would get 56% of the vote to Landon's 44%. His sample was apparently much more representative of the actual voting populace. The Gallup Organization went on to become one of the leading polling companies.

Selecting a sample to represent the population fairly is more difficult than it sounds. Polls or surveys most often fail because they use a sampling method that tends to over- or underrepresent parts of the population. The method may overlook subgroups that are harder to find (such as the homeless or those who use only cell phones) or favor others (such as Internet users who like to respond to online surveys). Sampling methods that, by their nature, tend to over- or underemphasize some characteristics of the population are said to be **biased**. Bias is the bane of sampling—the one thing above all to avoid. Conclusions based on samples drawn with biased methods are inherently flawed. There is usually no way to fix bias after the sample is drawn and no way to salvage useful information from it.

Here's a famous example of a really dismal failure. By the beginning of the 20th century, it was common for newspapers to ask readers to return "straw" ballots on a variety of topics. (Today's Internet surveys are the same idea, gone electronic.) The earliest known example of such a straw vote in the United States dates back to 1824.

During the period from 1916 to 1936, the magazine *Literary Digest* regularly surveyed public opinion and forecast election results correctly. During the 1936 presidential campaign between Alf Landon and Franklin Delano Roosevelt, it mailed more than 10 million ballots and got back an astonishing 2.4 million. (Polls were still a relatively novel idea, and many people thought it was important to send back their opinions.) The results were clear: Alf Landon would be the next president by a landslide, 57% to 43%. You remember President Landon? No? In fact, Landon carried only two states. Roosevelt won, 62% to 37%, and, perhaps coincidentally, the *Digest* went bankrupt soon afterward.

What went wrong? One problem was that the *Digest's* sample wasn't representative. Where would *you* find 10 million names and addresses to sample? The *Digest* used the phone book, as many surveys do.<sup>3</sup> But in 1936, at the height of the Great Depression, telephones were a real luxury, so they sampled more rich than poor voters. The campaign of 1936 focused on the economy, and those who were less well off were more likely to vote for the Democrat. So the *Digest's* sample was hopelessly biased.

How do modern polls get their samples to *represent* the entire population? You might think that they'd handpick individuals to sample with care and precision.

**A S** **Video: The *Literary Digest* Poll and the Election of 1936.** Hear the story of one of the most famous polling failures in history.

<sup>3</sup> Today phone numbers are computer-generated to make sure that unlisted numbers are included. But even now, cell phones and VOIP Internet phones are often not included.

But in fact, they do something quite different: They select individuals to sample *at random*. The importance of deliberately using randomness is one of the great insights of Statistics.

## Idea 2: Randomize



Think back to the soup sample. Suppose you add some salt to the pot. If you sample it from the top before stirring, you'll get the misleading idea that the whole pot is salty. If you sample from the bottom, you'll get an equally misleading idea that the whole pot is bland. By stirring, you *randomize* the amount of salt throughout the pot, making each taste more typical of the whole pot.

Not only does randomization protect you against factors that you know are in the data, it can also help protect against factors that you didn't even know were there. Suppose, while you weren't looking, a friend added a handful of peas to the soup. If they're down at the bottom of the pot, and you don't randomize the soup by stirring, your test spoonful won't have any peas. By stirring in the salt, you *also* randomize the peas throughout the pot, making your sample taste more typical of the overall pot *even though you didn't know the peas were there*. So randomizing protects us even in this case.

How do we "stir" people in a survey? We select them at random. **Randomizing** protects us from the influences of *all* the features of our population by making sure that, *on average*, the sample looks like the rest of the population.

**AS** **Activity: Sampling from Some Real Populations.** Draw random samples to see how closely they resemble each other and the population.

**Why not match the sample to the population?** Rather than randomizing, we could try to design our sample so that the people we choose are typical in terms of every characteristic we can think of. We might want the income levels of those we sample to match the population. How about age? Political affiliation? Marital status? Having children? Living in the suburbs? We can't possibly think of all the things that might be important. Even if we could, we wouldn't be able to match our sample to the population for all these characteristics.

### FOR EXAMPLE

#### Is a random sample representative?

Here are summary statistics comparing two samples of 8000 drawn at random from a company's database of 3.5 million customers:

Age (yr)	White (%)	Female (%)	# of Children	Income Bracket (1–7)	Wealth Bracket (1–9)	Homeowner? (% Yes)
61.4	85.12	56.2	1.54	3.91	5.29	71.36
61.2	84.44	56.4	1.51	3.88	5.33	72.30

**Question:** Do you think these samples are representative of the population? Explain.

The two samples look very similar with respect to these seven variables. It appears that randomizing has automatically matched them pretty closely. We can reasonably assume that since the two samples don't differ too much from each other, they don't differ much from the rest of the population either.

## Idea 3: It's the Sample Size

How large a random sample do we need for the sample to be reasonably representative of the population? Most people think that we need a large percentage, or *fraction*, of the population, but it turns out that what matters is the

**AS** **Activity: Does the Population Size Matter?** Here's the narrated version of this important idea about sampling.



A friend who knows that you are taking Statistics asks your advice on her study. What can you possibly say that will be helpful? Just say, "If you could just get a larger sample, it would probably improve your study." Even though a larger sample might not be worth the cost, it will almost always make the results more precise.

#### TI-*n*spire

**Populations and Samples.** How well can a sample reveal the population's shape, center, and spread? Explore what happens as you change the sample size.

number of individuals *in the sample*, not the size of the population. A random sample of 100 students in a college represents the student body just about as well as a random sample of 100 voters represents the entire electorate of the United States. This is the *third* idea and probably the most surprising one in designing surveys.

How can it be that only the size of the sample, and not the population, matters? Well, let's return one last time to that pot of soup. If you're cooking for a banquet rather than just for a few people, your pot will be bigger, but do you need a bigger spoon to decide how the soup tastes? Of course not. The same-size spoonful is probably enough to make a decision about the entire pot, no matter how large the pot. **The fraction of the population that you've sampled doesn't matter.<sup>4</sup> It's the sample size itself that's important.**

How big a sample do you need? That depends on what you're estimating. To get an idea of what's really in the soup, you'll need a large enough taste to get a *representative* sample from the pot. For a survey that tries to find the proportion of the population falling into a category, you'll usually need several hundred respondents to say anything precise enough to be useful.<sup>5</sup>

**What do the pollsters do?** How do professional polling agencies do their work? The most common polling method today is to contact respondents by telephone. Computers generate random telephone numbers, so pollsters can even call some people with unlisted phone numbers. The person who answers the phone is invited to respond to the survey—if that person qualifies. (For example, only if it's an adult who lives at that address.) If the person answering doesn't qualify, the caller will ask for an appropriate alternative. In phrasing questions, pollsters often list alternative responses (such as candidates' names) in different orders to avoid biases that might favor the first name on the list.

Do these methods work? The Pew Research Center for the People and the Press, reporting on one survey, says that

*Across five days of interviewing, surveys today are able to make some kind of contact with the vast majority of households (76%), and there is no decline in this contact rate over the past seven years. But because of busy schedules, skepticism and outright refusals, interviews were completed in just 38% of households that were reached using standard polling procedures.*

Nevertheless, studies indicate that those actually sampled can give a good snapshot of larger populations from which the surveyed households were drawn.

## Does a Census Make Sense?

**AS** **Video: Frito-Lay Sampling for Quality.** How does a potato chip manufacturer make sure to cook only the best potatoes?

Why bother determining the right sample size? **Wouldn't it be better to just include everyone and "sample" the entire population? Such a special sample is called a census.** Although a census would appear to provide the best possible information about the population, there are a number of reasons why it might not.

First, it can be difficult to complete a census. Some individuals in the population will be hard (and expensive) to locate. Or a census might just be impractical. If you were a taste tester for the Hostess™ Company, you probably wouldn't want to census *all* the Twinkies on the production line. Not only might this be life-endangering, but you wouldn't have any left to sell.

<sup>4</sup> Well, that's not exactly true. If the population is small enough and the sample is more than 10% of the whole population, it *can* matter. It doesn't matter whenever, as usual, our sample is a very small fraction of the population.

<sup>5</sup> Chapter 19 gives the details behind this statement and shows how to decide on a sample size for a survey.

Second, populations rarely stand still. In populations of people, babies are born and folks die or leave the country. In opinion surveys, events may cause a shift in opinion during the survey. A census takes longer to complete and the population changes while you work. A sample surveyed in just a few days may give more accurate information.

Third, taking a census can be more complex than sampling. For example, the U.S. Census records too many college students. Many are counted once with their families and are then counted a second time in a report filed by their schools.

**The undercount.** It's particularly difficult to compile a complete census of a population as large, complex, and spread out as the U.S. population. The U.S. Census is known to miss some residents. On occasion, the undercount has been striking. For example, there have been blocks in inner cities in which the number of residents recorded by the Census was smaller than the number of electric meters for which bills were being paid. What makes the problem particularly important is that some groups have a higher probability of being missed than others—undocumented immigrants, the homeless, the poor. The Census Bureau proposed the use of random sampling to estimate the number of residents missed by the ordinary census. Unfortunately, the resulting debate has become more political than statistical.

## Populations and Parameters

Any quantity that we calculate from data could be called a “statistic.” But in practice, we usually use a statistic to estimate a population parameter.

**AS** **Activity: Statistics and Parameters.** Explore the difference between statistics and parameters.

**Remember: Population model parameters are not just unknown—usually they are *unknowable*. We have to settle for sample statistics.**

A study found that teens were less likely to “buckle up.” The National Center for Chronic Disease Prevention and Health Promotion reports that 21.7% of U.S. teens never or rarely wear seatbelts. We’re sure they didn’t take a census, so what *does* the 21.7% mean? We can’t know what percentage of teenagers wear seatbelts. Reality is just too complex. But we can simplify the question by building a model.

Models use mathematics to represent reality. Parameters are the key numbers in those models. A parameter used in a model for a population is sometimes called (redundantly) a **population parameter**.

But let’s not forget about the data. We use summaries of the data to estimate the population parameters. As we know, any summary found from the data is a **statistic**. Sometimes you’ll see the (also redundant) term **sample statistic**.<sup>6</sup>

We’ve already met two parameters in Chapter 6: the mean,  $\mu$ , and the standard deviation,  $\sigma$ . We’ll try to keep denoting population model parameters with Greek letters and the corresponding statistics with Latin letters. Usually, but not always, the letter used for the statistic and the parameter correspond in a natural way. So the standard deviation of the data is  $s$ , and the corresponding parameter is  $\sigma$  (Greek for  $s$ ). In Chapter 7, we used  $r$  to denote the sample correlation. The corresponding correlation in a model for the population would be called  $\rho$  (rho). In Chapter 8,  $b_1$  represented the slope of a linear regression estimated from the data. But when we think about a (linear) *model* for the population, we denote the slope parameter  $\beta_1$  (beta).

Get the pattern? Good. Now it breaks down. We denote the mean of a population model with  $\mu$  (because  $\mu$  is the Greek letter for  $m$ ). It might make sense to denote the sample mean with  $m$ , but long-standing convention is to put a bar over anything when we average it, so we write  $\bar{y}$ . What about proportions? Suppose we want to talk about the proportion of teens who don’t wear seatbelts. If we use  $p$  to denote the proportion from the data, what is the corresponding model parameter? By all rights it should be  $\pi$ . But statements like  $\pi = 0.25$  might be confusing because  $\pi$  has been equal to 3.1415926 . . . for so long, and it’s worked so *well*. So, once again we violate the rule. We’ll use  $p$  for the population model

<sup>6</sup> Where else besides a sample *could* a statistic come from?

parameter and  $\hat{p}$  for the proportion from the data (since, like  $\hat{y}$  in regression, it's an estimated value).

Here's a table summarizing the notation:

### NOTATION ALERT:

This entire table is a notation alert.

Name	Statistic	Parameter
Mean	$\bar{y}$	$\mu$ (mu, pronounced "meeoo," not "moo")
Standard deviation	$s$	$\sigma$ (sigma)
Correlation	$r$	$\rho$ (rho)
Regression coefficient	$b$	$\beta$ (beta, pronounced "baytah" <sup>7</sup> )
Proportion	$\hat{p}$	$p$ (pronounced "pee" <sup>8</sup> )

We draw samples because we can't work with the entire population, but we want the statistics we compute from a sample to reflect the corresponding parameters accurately. A sample that does this is said to be **representative**. A biased sampling methodology tends to over- or underestimate the parameter of interest.



### JUST CHECKING

1. Various claims are often made for surveys. Why is each of the following claims not correct?
  - a) It is always better to take a census than to draw a sample.
  - b) Stopping students on their way out of the cafeteria is a good way to sample if we want to know about the quality of the food there.
  - c) We drew a sample of 100 from the 3000 students in a school. To get the same level of precision for a town of 30,000 residents, we'll need a sample of 1000.
  - d) A poll taken at a statistics support Web site garnered 12,357 responses. The majority said they enjoy doing statistics homework. With a sample size that large, we can be pretty sure that most Statistics students feel this way, too.
  - e) The true percentage of all Statistics students who enjoy the homework is called a "population statistic."

## Simple Random Samples

How would you select a representative sample? Most people would say that every individual in the population should have an equal chance to be selected, and certainly that seems fair. But it's not sufficient. There are many ways to give everyone an equal chance that still wouldn't give a representative sample. Consider, for example, a school that has equal numbers of males and females. We could sample like this: Flip a coin. If it comes up heads, select 100 female students at random. If it comes up tails, select 100 males at random. Everyone has an equal chance of selection, but every sample is of only a single sex—hardly representative.

We need to do better. Suppose we insist that every possible *sample* of the size we plan to draw has an equal chance to be selected. This ensures that situations like the one just described are not likely to occur and still guarantees that each person has an equal chance of being selected. What's different is that with this method, each *combination* of people has an equal chance of being selected as well. A sample drawn in this way is called a **Simple Random Sample**, usually abbreviated **SRS**. An SRS is the standard against which we measure other sampling methods, and the sampling method on which the theory of working with sampled data is based.

To select a sample at random, we first need to define where the sample will come from. The **sampling frame** is a list of individuals from which the sample is drawn.

<sup>7</sup> If you're from the United States. If you're British or Canadian, it's "beetah."

<sup>8</sup> Just in case you weren't sure.

For example, to draw a random sample of students at a college, we might obtain a list of all registered full-time students and sample from that list. In defining the sampling frame, we must deal with the details of defining the population. Are part-time students included? How about those who are attending school elsewhere and transferring credits back to the college?

Once we have a sampling frame, the easiest way to choose an SRS is to assign a random number to each individual in the sampling frame. We then select only those whose random numbers satisfy some rule.<sup>9</sup> Let's look at some ways to do this.

## FOR EXAMPLE

### Using random numbers to get an SRS

There are 80 students enrolled in an introductory Statistics class; you are to select a sample of 5.

**Question:** How can you select an SRS of 5 students using these random digits found on the Internet: 05166 29305 77482?

First I'll number the students from 00 to 79. Taking the random numbers two digits at a time gives me 05, 16, 62, 93, 05, 77, and 48. I'll ignore 93 because the students were numbered only up to 79. And, so as not to pick the same person twice, I'll skip the repeated number 05. My simple random sample consists of students with the numbers 05, 16, 62, 77, and 48.

### Error Okay, Bias Bad!

Sampling variability is sometimes referred to as *sampling error*, making it sound like it's some kind of mistake. It's not. We understand that samples will vary, so "sampling error" is to be expected. It's *bias* we must strive to avoid. Bias means our sampling method distorts our view of the population, and that will surely lead to mistakes.

- ▶ We can be more efficient when we're choosing a larger sample from a sampling frame stored in a data file. First we assign a random number with several digits (say, from 0 to 10,000) to each individual. Then we arrange the random numbers in numerical order, keeping each name with its number. Choosing the first  $n$  names from this re-arranged list will give us a random sample of that size.
- ▶ Often the sampling frame is so large that it would be too tedious to number everyone consecutively. If our intended sample size is approximately 10% of the sampling frame, we can assign each individual a single random digit 0 to 9. Then we select only those with a specific random digit, say, 5.

Samples drawn at random generally differ one from another. Each draw of random numbers selects *different* people for our sample. These differences lead to different values for the variables we measure. We call these sample-to-sample differences **sampling variability**. Surprisingly, sampling variability isn't a problem; it's an opportunity. In future chapters we'll investigate what the variation in a sample can tell us about its population.

## Stratified Sampling

Simple random sampling is not the only fair way to sample. More complicated designs may save time or money or help avoid sampling problems. All statistical sampling designs have in common the idea that chance, rather than human choice, is used to select the sample.

Designs that are used to sample from large populations—especially populations residing across large areas—are often more complicated than simple random samples. Sometimes the population is first sliced into homogeneous groups, called **strata**, before the sample is selected. Then simple random sampling is used within each stratum before the results are combined. This common sampling design is called **stratified random sampling**.

Why would we want to complicate things? Here's an example. Suppose we want to learn how students feel about funding for the football team at a large

<sup>9</sup> Chapter 11 presented ways of finding and working with random numbers.



university. The campus is 60% men and 40% women, and we suspect that men and women have different views on the funding. If we use simple random sampling to select 100 people for the survey, we could end up with 70 men and 30 women or 35 men and 65 women. Our resulting estimates of the level of support for the football funding could vary widely. To help reduce this sampling variability, we can decide to force a representative balance, selecting 60 men at random and 40 women at random. This would guarantee that the proportions of men and women within our sample match the proportions in the population, and that should make such samples more accurate in representing population opinion.

You can imagine the importance of stratifying by race, income, age, and other characteristics, depending on the questions in the survey. Samples taken within a stratum vary less, so our estimates can be more precise. **This reduced sampling variability is the most important benefit of stratifying.**

Stratified sampling can also help us notice important differences among groups. As we saw in Chapter 3, if we unthinkingly combine group data, we risk reaching the wrong conclusion, becoming victims of Simpson's paradox.

### FOR EXAMPLE

#### Stratifying the sample

**Recap:** You're trying to find out what freshmen think of the food served on campus. Food Services believes that men and women typically have different opinions about the importance of the salad bar.

**Question:** How should you adjust your sampling strategy to allow for this difference?

*I will stratify my sample by drawing an SRS of men and a separate SRS of women—assuming that the data from the registrar include information about each person's sex.*

## Cluster and Multistage Sampling

Suppose we wanted to assess the reading level of this textbook based on the length of the sentences. Simple random sampling could be awkward; we'd have to number each sentence, then find, for example, the 576th sentence or the 2482nd sentence, and so on. Doesn't sound like much fun, does it?

It would be much easier to pick a few *pages* at random and count the lengths of the sentences on those pages. That works if we believe that each page is representative of the entire book in terms of reading level. **Splitting the population into representative clusters can make sampling more practical.** Then we could simply select one or a few clusters at random and perform a census within each of them. **This sampling design is called cluster sampling.** If each cluster represents the full population fairly, cluster sampling will be unbiased.

### FOR EXAMPLE

#### Cluster sampling

**Recap:** In trying to find out what freshmen think about the food served on campus, you've considered both an SRS and a stratified sample. Now you have run into a problem: It's simply too difficult and time consuming to track down the individuals whose names were chosen for your sample. Fortunately, freshmen at your school are all housed in 10 freshman dorms.

**Questions:** How could you use this fact to draw a cluster sample? How might that alleviate the problem? What concerns do you have?

*To draw a cluster sample, I would select one or two dorms at random and then try to contact everyone in each selected dorm. I could save time by simply knocking on doors on a given evening and interviewing people. I'd have to assume that freshmen were assigned to dorms pretty much at random and that the people I'm able to contact are representative of everyone in the dorm.*

What's the difference between cluster sampling and stratified sampling? We stratify to ensure that our sample represents different groups in the population, and we sample randomly within each stratum. Strata are internally homogeneous, but differ from one another. By contrast, clusters are internally heterogeneous, each resembling the overall population. We select clusters to make sampling more practical or affordable.



**Stratified vs. cluster sampling.** Boston cream pie consists of a layer of yellow cake, a layer of pastry creme, another cake layer, and then a chocolate frosting. Suppose you are a professional taster (yes, there really are such people) whose job is to check your company's pies for quality. You'd need to eat small samples of randomly selected pies, tasting all three components: the cake, the creme, and the frosting.

One approach is to cut a thin vertical slice out of the pie. Such a slice will be a lot like the entire pie, so by eating that slice, you'll learn about the whole pie. This vertical slice containing all the different ingredients in the pie would be a *cluster* sample.

Another approach is to sample in *strata*: Select some tastes of the cake at random, some tastes of creme at random, and some bits of frosting at random. You'll end up with a reliable judgment of the pie's quality.

Many populations you might want to learn about are like this Boston cream pie. You can think of the subpopulations of interest as horizontal strata, like the layers of pie. Cluster samples slice vertically across the layers to obtain clusters, each of which is representative of the entire population. Stratified samples represent the population by drawing some from each layer, reducing variability in the results that could arise because of the differences among the layers.

### Strata or Clusters?

We may split a population into strata or clusters. What's the difference? We create strata by dividing the population into groups of similar individuals so that each stratum is different from the others. By contrast, since clusters each represent the entire population, they all look pretty much alike.

Sometimes we use a variety of sampling methods together. In trying to assess the reading level of this book, we might worry that it starts out easy and then gets harder as the concepts become more difficult. If so, we'd want to avoid samples that selected heavily from early or from late chapters. To guarantee a fair mix of chapters, we could randomly choose one chapter from each of the seven parts of the book and then randomly select a few pages from each of those chapters. If, altogether, that made too many sentences, we might select a few sentences at random from each of the chosen pages. So, what is our sampling strategy? First we stratify by the part of the book and randomly choose a chapter to represent each stratum. Within each selected chapter, we choose pages as clusters. Finally, we consider an SRS of sentences within each cluster. **Sampling schemes that combine several methods are called multistage samples.** Most surveys conducted by professional polling organizations use some combination of stratified and cluster sampling as well as simple random samples.

### FOR EXAMPLE

#### Multistage sampling

**Recap:** Having learned that freshmen are housed in separate dorms allowed you to sample their attitudes about the campus food by going to dorms chosen at random, but you're still concerned about possible differences in opinions between men and women. It turns out that these freshmen dorms house the sexes on alternate floors.

**Question:** How can you design a sampling plan that uses this fact to your advantage?

Now I can stratify my sample by sex. I would first choose one or two dorms at random and then select some dorm floors at random from among those that house men and, separately, from among those that house women. I could then treat each floor as a cluster and interview everyone on that floor.

## Systematic Samples

Some samples select individuals systematically. For example, you might survey every 10th person on an alphabetical list of students. To make it random, you still must start the systematic selection from a randomly selected individual. When the order of the list is not associated in any way with the responses sought, **systematic sampling** can give a representative sample. Systematic sampling can be much less expensive than true random sampling. When you use a systematic sample, you should justify the assumption that the systematic method is not associated with any of the measured variables.

Think about the reading-level sampling example again. Suppose we have chosen a chapter of the book at random, then three pages at random from that chapter, and now we want to select a sample of 10 sentences from the 73 sentences found on those pages. Instead of numbering each sentence so we can pick a simple random sample, it would be easier to sample systematically. A quick calculation shows  $73/10 = 7.3$ , so we can get our sample by just picking every seventh sentence on the page. But where should you start? At random, of course. We've accounted for  $10 \times 7 = 70$  of the sentences, so we'll throw the extra 3 into the starting group and choose a sentence at random from the first 10. Then we pick every seventh sentence after that and record its length.



### JUST CHECKING

2. We need to survey a random sample of the 300 passengers on a flight from San Francisco to Tokyo. Name each sampling method described below.
  - a) Pick every 10th passenger as people board the plane.
  - b) From the boarding list, randomly choose 5 people flying first class and 25 of the other passengers.
  - c) Randomly generate 30 seat numbers and survey the passengers who sit there.
  - d) Randomly select a seat position (right window, right center, right aisle, etc.) and survey all the passengers sitting in those seats.

### STEP-BY-STEP EXAMPLE

### Sampling

The assignment says, "Conduct your own sample survey to find out how many hours per week students at your school spend watching TV during the school year." Let's see how we might do this step by step. (Remember, though—actually collecting the data from your sample can be difficult and time consuming.)

**Question:** How would you design this survey?



**Plan** State what you want to know.

**Population and Parameter** Identify the W's of the study. The *Why* determines the population and the associated sampling frame. The *What* identifies the parameter of interest and the variables measured. The *Who* is the sample we actually draw. The *How*, *When*, and *Where* are given by the sampling plan.

*I wanted to design a study to find out how many hours of TV students at my school watch.*

*The population studied was students at our school. I obtained a list of all students currently enrolled and used it as the sampling frame. The parameter of interest was the number of TV hours watched per week during the school year, which I attempted to measure by asking students how much TV they watched during the previous week.*

Often, thinking about the *Why* will help us see whether the sampling frame and plan are adequate to learn about the population.

**Sampling Plan** Specify the sampling method and the sample size,  $n$ . Specify how the sample was actually drawn. What is the sampling frame? How was the randomization performed?

A good description should be complete enough to allow someone to replicate the procedure, drawing another sample from the same population in the same manner.

*I decided against stratifying by class or sex because I didn't think TV watching would differ much between males and females or across classes. I selected a simple random sample of students from the list. I obtained an alphabetical list of students, assigned each a random digit between 0 and 9, and then selected all students who were assigned a "4." This method generated a sample of 212 students from the population of 2133 students.*



**Sampling Practice** Specify *When*, *Where*, and *How* the sampling was performed. Specify any other details of your survey, such as how respondents were contacted, what incentives were offered to encourage them to respond, how nonrespondents were treated, and so on.

*The survey was taken over the period Oct. 15 to Oct. 25. Surveys were sent to selected students by e-mail, with the request that they respond by e-mail as well. Students who could not be reached by e-mail were handed the survey in person.*



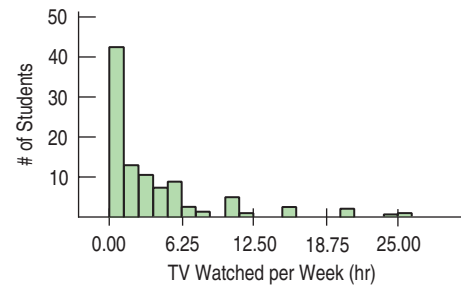
**Summary and Conclusion** This report should include a discussion of all the elements. In addition, it's good practice to discuss any special circumstances. Professional polling organizations report the *When* of their samples but will also note, for example, any important news that might have changed respondents' opinions during the sampling process. In this survey, perhaps, a major news story or sporting event might change students' TV viewing behavior.

The question you ask also matters. It's better to be specific ("How many hours did you watch TV last week?") than to ask a general question ("How many hours of TV do you usually watch in a week?").

*During the period Oct. 15 to Oct. 25, 212 students were randomly selected, using a simple random sample from a list of all students currently enrolled. The survey they received asked the following question: "How many hours did you spend watching television last week?"*

*Of the 212 students surveyed, 110 responded. It's possible that the nonrespondents differ in the number of TV hours watched from those who responded, but I was unable to follow up on them due to limited time and funds. The 110 respondents reported an average 3.62 hours of TV watching per week. The median was only 2 hours per week. A histogram of the data shows that the distribution is highly right-skewed, indicating that the median might be a more appropriate summary of the typical TV watching of the students.*

The report should show a display of the data, provide and interpret the statistics from the sample, and state the conclusions that you reached about the population.



Most of the students (90%) watch between 0 and 10 hours per week, while 30% reported watching less than 1 hour per week. A few watch much more. About 3% reported watching more than 20 hours per week.

## Defining the “Who”: You Can’t Always Get What You Want

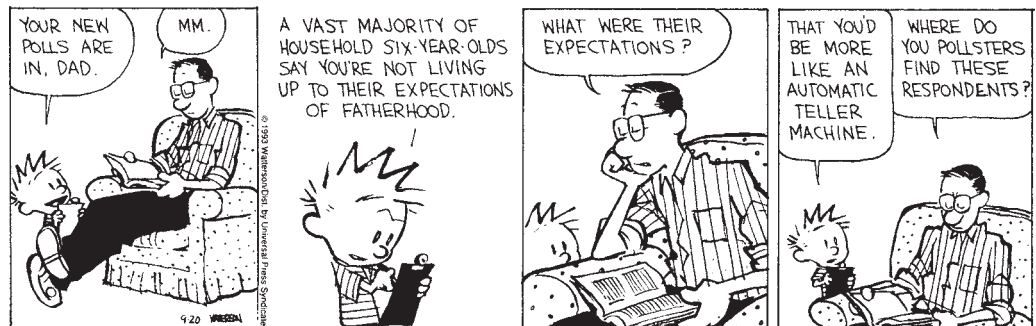
Before you start a survey, think first about the population you want to study. You may find that it’s not the well-defined group you thought it was. Who, exactly, is a student, for example? Even if the population seems well defined, it may not be a practical group from which to draw a sample. For example, election polls want to sample from all those who will vote in the next election—a population that is impossible to identify before Election Day.

The population is determined by the *Why* of the study. Unfortunately, the sample is just those we can reach to obtain responses—the *Who* of the study. This difference could undermine even a well-designed study.

Next, you must specify the sampling frame. (Do you have a list of students to sample from? How about a list of registered voters?) Usually, the sampling frame is not the group you *really* want to know about. (All those registered to vote are not equally likely to show up.) The sampling frame limits what your survey can find out.

Then there’s your target sample. These are the individuals for whom you *intend* to measure responses. You’re not likely to get responses from all of them. (“I know it’s dinnertime, but I’m sure you wouldn’t mind answering a few questions. It’ll only take 20 minutes or so. Oh, you’re busy?”) Nonresponse is a problem in many surveys.

Finally, there’s your sample—the actual respondents. These are the individuals about whom you *do* get data and can draw conclusions. Unfortunately, they might not be representative of the sampling frame or the population.



At each step, the group we can study may be constrained further. The *Who* keeps changing, and each constraint can introduce biases. A careful study should address the question of how well each group matches the population of interest. One of the main benefits of simple random sampling is that it never loses its sense of who's *Who*. The *Who* in an SRS is the population of interest from which we've drawn a representative sample. That's not always true for other kinds of samples.

## The Valid Survey

It isn't sufficient to just draw a sample and start asking questions. We'll want our survey to be *valid*. A valid survey yields the information we are seeking about the population we are interested in. Before setting out to survey, ask yourself:

- ▶ What do I want to know?
- ▶ Am I asking the right respondents?
- ▶ Am I asking the right questions?
- ▶ What would I do with the answers if I had them; would they address the things I want to know?

These questions may sound obvious, but there are a number of pitfalls to avoid.

*Know what you want to know.* Before considering a survey, understand what you hope to learn and about whom you hope to learn it. Far too often, people decide to perform a survey without any clear idea of what they hope to learn.

*Use the right frame.* A valid survey obtains responses from the appropriate respondents. Be sure you have a suitable *sampling frame*. Have you identified the population of interest and sampled from it appropriately? A company might survey customers who returned warranty registration cards, a readily available sampling frame. But if the company wants to know how to make their product more attractive, the most important population is the customers who rejected their product in favor of one from a competitor.

*Tune your instrument.* It is often tempting to ask questions you don't really need, but beware—longer questionnaires yield fewer responses and thus a greater chance of nonresponse bias.

*Ask specific rather than general questions.* People are not very good at estimating their typical behavior, so it is better to ask "How many hours did you sleep last night?" than "How much do you usually sleep?" Sure, some responses will include some unusual events (My dog was sick; I was up all night.), but overall you'll get better data.

*Ask for quantitative results when possible.* "How many magazines did you read last week?" is better than "How much do you read: A lot, A moderate amount, A little, or None at all?"

*Be careful in phrasing questions.* A respondent may not understand the question—or may understand the question differently than the researcher intended it. ("Does anyone in your family belong to a union?" Do you mean just me, my spouse, and my children? Or does "family" include my father, my siblings, and my second cousin once removed? What about my grandfather, who is staying with us? I think he once belonged to the Autoworkers Union.) Respondents are unlikely (or may not have the opportunity) to ask for clarification. A question like "Do you approve of the recent actions of the Secretary of Labor?" is likely not to measure what you want if many re-

spondents don't know who the Secretary of Labor is or what actions he or she recently made.

Respondents may even lie or shade their responses if they feel embarrassed by the question ("Did you have too much to drink last night?"), are intimidated or insulted by the question ("Could you understand our new *Instructions for Dummies* manual, or was it too difficult for you?"), or if they want to avoid offending the interviewer ("Would you hire a man with a tattoo?" asked by a tattooed interviewer). Also, be careful to avoid phrases that have double or regional meanings. "How often do you go to town?" might be interpreted differently by different people and cultures.

*Even subtle differences in phrasing can make a difference.* In January 2006, the *New York Times* asked half of the 1229 U.S. adults in their sample the following question:

*After 9/11, President Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants, saying this was necessary to reduce the threat of terrorism. Do you approve or disapprove of this?*

They found that 53% of respondents approved. But when they asked the other half of their sample a question with only slightly different phrasing,

*After 9/11, George W. Bush authorized government wiretaps on some phone calls in the U.S. without getting court warrants. Do you approve or disapprove of this?*

only 46% approved.

*Be careful in phrasing answers.* It's often a good idea to offer choices rather than inviting a free response. Open-ended answers can be difficult to analyze. "How did you like the movie?" may start an interesting debate, but it may be better to give a range of possible responses. Be sure to phrase them in a neutral way. When asking "Do you support higher school taxes?" positive responses could be worded "Yes," "Yes, it is important for our children," or "Yes, our future depends on it." But those are not equivalent answers.

#### THE WIZARD OF ID



The best way to protect a survey from such unanticipated measurement errors is to perform a pilot survey. A **pilot** is a trial run of the survey you eventually plan to give to a larger group, using a draft of your survey questions administered to a small sample drawn from the same sampling frame you intend to use. By analyzing the results from this smaller survey, you can often discover ways to improve your instrument.

## WHAT CAN GO WRONG?—OR, HOW TO SAMPLE BADLY

Bad sample designs yield worthless data. Many of the most convenient forms of sampling can be seriously biased. And there is no way to correct for the bias from a bad sample. So it's wise to pay attention to sample design—and to beware of reports based on poor samples.

### SAMPLE BADLY WITH VOLUNTEERS

One of the most common dangerous sampling methods is a voluntary response sample. In a **voluntary response sample**, a large group of individuals is invited to respond, and all who do respond are counted. This method is used by call-in shows, 900 numbers, Internet polls, and letters written to members of Congress. Voluntary response samples are almost always biased, and so conclusions drawn from them are almost always wrong.

It's often hard to define the sampling frame of a voluntary response study. Practically, the frames are groups such as Internet users who frequent a particular Web site or those who happen to be watching a particular TV show at the moment. But those sampling frames don't correspond to interesting populations.

Even within the sampling frame, voluntary response samples are often biased toward those with strong opinions or those who are strongly motivated. People with very negative opinions tend to respond more often than those with equally strong positive opinions. The sample is not representative, even though every individual in the population may have been offered the chance to respond. The resulting **voluntary response bias** invalidates the survey.

**AS** **Activity: Sources of Sampling Bias.** Here's a narrated exploration of sampling bias.

**If you had it to do over again, would you have children?** Ann Landers, the advice columnist, asked parents this question. The overwhelming majority—70% of the more than 10,000 people who wrote in—said no, kids weren't worth it. A more carefully designed survey later showed that about 90% of parents actually are happy with their decision to have children. What accounts for the striking difference in these two results? What parents do you think are most likely to respond to the original question?

### FOR EXAMPLE

#### Bias in sampling

**Recap:** You're trying to find out what freshmen think of the food served on campus, and have thought of a variety of sampling methods, all time consuming. A friend suggests that you set up a "Tell Us What You Think" Web site and invite freshmen to visit the site to complete a questionnaire.

**Question:** What's wrong with this idea?

*Letting each freshman decide whether to participate makes this a voluntary response survey. Students who were dissatisfied might be more likely to go to the Web site to record their complaints, and this could give me a biased view of the opinions of all freshmen.*

### SAMPLE BADLY, BUT CONVENIENTLY

Another sampling method that doesn't work is convenience sampling. As the name suggests, in **convenience sampling** we simply include the individuals who are convenient for us to sample. Unfortunately, this group may not be representative of the population. A recent survey of 437 potential home buyers in Orange County, California, found, among other things, that

Do you use the Internet?

Click here  for yes

Click here  for no



Internet convenience surveys are worthless. As voluntary response surveys, they have no well-defined sampling frame (all those who use the Internet and visit their site?) and thus report no useful information. Do not believe them.

*All but 2 percent of the buyers have at least one computer at home, and 62 percent have two or more. Of those with a computer, 99 percent are connected to the Internet (Jennifer Hieger, "Portrait of Homebuyer Household: 2 Kids and a PC," Orange County Register, 27 July 2001).*

Later in the article, we learn that the survey was conducted via the Internet! That was a convenient way to collect data and surely easier than drawing a simple random sample, but perhaps home builders shouldn't conclude from this study that *every* family has a computer and an Internet connection.

Many surveys conducted at shopping malls suffer from the same problem. People in shopping malls are not necessarily representative of the population of interest. Mall shoppers tend to be more affluent and include a larger percentage of teenagers and retirees than the population at large. To make matters worse, survey interviewers tend to select individuals who look "safe," or easy to interview.

## FOR EXAMPLE

### Bias in sampling

**Recap:** To try to gauge freshman opinion about the food served on campus, Food Services suggests that you just stand outside a school cafeteria at lunchtime and stop people to ask them questions.

**Questions:** What's wrong with this sampling strategy?

*This would be a convenience sample, and it's likely to be biased. I would miss people who use the cafeteria for dinner, but not for lunch, and I'd never hear from anyone who hates the food so much that they have stopped coming to the school cafeterias.*

## SAMPLE FROM A BAD SAMPLING FRAME

An SRS from an incomplete sampling frame introduces bias because the individuals included may differ from the ones not in the frame. People in prison, homeless people, students, and long-term travelers are all likely to be missed. In telephone surveys, people who have only cell phones or who use VOIP Internet phones are often missing from the sampling frame.

## UNDERCOVERAGE

Many survey designs suffer from **undercoverage**, in which some portion of the population is not sampled at all or has a smaller representation in the sample than it has in the population. Undercoverage can arise for a number of reasons, but it's always a potential source of bias.

Telephone surveys are usually conducted when you are likely to be home, interrupting your dinner. If you eat out often, you may be less likely to be surveyed, a possible source of undercoverage.

## WHAT <sup>else</sup> CAN GO WRONG?

- ▶ **Watch out for nonrespondents.** A common and serious potential source of bias for most surveys is **nonresponse bias**. No survey succeeds in getting responses from everyone. The problem is that those who don't respond may differ from those who do. And they may differ on just the variables we care about. The lack of response will

(continued)

bias the results. Rather than sending out a large number of surveys for which the response rate will be low, it is often better to design a smaller randomized survey for which you have the resources to ensure a high response rate. One of the problems with nonresponse bias is that it's usually impossible to tell what the nonrespondents might have said.

**AS** **Video: Biased Question Wording.** Watch a hapless interviewer make every mistake in the book.

### A Short Survey

Given the fact that those who understand Statistics are smarter and better looking than those who don't, don't you think it is important to take a course in Statistics?

**AS** **Activity: Can a Large Sample Protect Against Bias?** Explore how we can learn about the population from large or repeated samples.

A researcher distributed a survey to an organization before some economizing changes were made. She asked how people felt about a proposed cutback in secretarial and administrative support on a seven-point scale from Very Happy to Very Unhappy.

But virtually all respondents were very unhappy about the cutbacks, so the results weren't particularly useful. If she had pretested the question, she might have chosen a scale that ran from Unhappy to Outraged.

**Remember the *Literary Digest* Survey?** It turns out that they were wrong on two counts. First, their list of 10 million people was not representative. There was a selection bias in their sampling frame. There was also a nonresponse bias. We know this because the *Digest* also surveyed a *systematic* sample in Chicago, sending the same question used in the larger survey to every third registered voter. They *still* got a result in favor of Landon, even though Chicago voted overwhelmingly for Roosevelt in the election. This suggests that the Roosevelt supporters were less likely to respond to the *Digest* survey. There's a modern version of this problem: It's been suggested that those who screen their calls with caller ID or an answering machine, and so might not talk to a pollster, may differ in wealth or political views from those who just answer the phone.

- ▶ **Work hard to avoid influencing responses.** **Response bias**<sup>10</sup> refers to anything in the survey design that influences the responses. Response biases include the tendency of respondents to tailor their responses to try to please the interviewer, the natural unwillingness of respondents to reveal personal facts or admit to illegal or unapproved behavior and the ways in which the wording of the questions can influence responses.

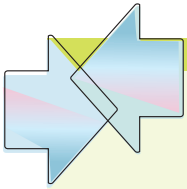
## HOW TO THINK ABOUT BIASES

- ▶ **Look for biases in any survey you encounter.** If you design one of your own, ask someone else to help look for biases that may not be obvious to you. And do this *before* you collect your data. **There's no way to recover from a biased sampling method or a survey that asks biased questions.** Sorry, it just can't be done.

A bigger sample size for a biased study just gives you a bigger useless study. A really big sample gives you a really big useless study. (Think of the 2.4 million *Literary Digest* responses.)

- ▶ **Spend your time and resources reducing biases.** No other use of resources is as worthwhile as reducing the biases.
- ▶ **If you can, pilot-test your survey.** Administer the survey in the exact form that you intend to use it to a small sample drawn from the population you intend to sample. Look for misunderstandings, misinterpretation, confusion, or other possible biases. Then refine your survey instrument.
- ▶ **Always report your sampling methods in detail.** Others may be able to detect biases where you did not expect to find them.

<sup>10</sup> Response bias is not the opposite of nonresponse bias. (We don't make these terms up; we just try to explain them.)



## CONNECTIONS

With this chapter, we take our first formal steps to relate our sample data to a larger population. Some of these ideas have been lurking in the background as we sought patterns and summaries for data. Even when we only worked with the data at hand, we often thought about implications for a larger population of individuals.

Notice the ongoing central importance of models. We've seen models in several ways in previous chapters. Here we recognize the value of a model for a population. The parameters of such a model are values we will often want to estimate using statistics such as those we've been calculating. The connections to summary statistics for center, spread, correlation, and slope are obvious.

We now have a specific application for random numbers. The idea of applying randomness deliberately showed up in Chapter 11 for simulation. Now we need randomization to get good-quality data from the real world.

## WHAT HAVE WE LEARNED?



We've learned that a representative sample can offer us important insights about populations. It's the size of the sample—and not its fraction of the larger population—that determines the precision of the statistics it yields.

We've learned several ways to draw samples, all based on the power of randomness to make them representative of the population of interest:

- ▶ A Simple Random Sample (SRS) is our standard. Every possible group of  $n$  individuals has an equal chance of being our sample. That's what makes it *simple*.
- ▶ Stratified samples can reduce sampling variability by identifying homogeneous subgroups and then randomly sampling within each.
- ▶ Cluster samples randomly select among heterogeneous subgroups that each resemble the population at large, making our sampling tasks more manageable.
- ▶ Systematic samples can work in some situations and are often the least expensive method of sampling. But we still want to start them randomly.
- ▶ Multistage samples combine several random sampling methods.

We've learned that bias can destroy our ability to gain insights from our sample:

- ▶ Nonresponse bias can arise when sampled individuals will not or cannot respond.
- ▶ Response bias arises when respondents' answers might be affected by external influences, such as question wording or interviewer behavior.

We've learned that bias can also arise from poor sampling methods:

- ▶ Voluntary response samples are almost always biased and should be avoided and distrusted.
- ▶ Convenience samples are likely to be flawed for similar reasons.
- ▶ Even with a reasonable design, sample frames may not be representative. Undercoverage occurs when individuals from a subgroup of the population are selected less often than they should be.

Finally, we've learned to look for biases in any survey we find and to be sure to report our methods whenever we perform a survey so that others can evaluate the fairness and accuracy of our results.

## Terms

**Population**  
**Sample**

268. The entire group of individuals or instances about whom we hope to learn.

268. A (representative) subset of a population, examined in hope of learning about the population.

Sample survey	269. A study that asks questions of a sample drawn from some population in the hope of learning something about the entire population. Polls taken to assess voter preferences are common sample surveys.
Bias	269. Any systematic failure of a sampling method to represent its population is bias. Biased sampling methods tend to over- or underestimate parameters. It is almost impossible to recover from bias, so efforts to avoid it are well spent. Common errors include <ul style="list-style-type: none"> <li>▶ relying on voluntary response.</li> <li>▶ undercoverage of the population.</li> <li>▶ nonresponse bias.</li> <li>▶ response bias.</li> </ul>
Randomization	270. The best defense against bias is randomization, in which each individual is given a fair, random chance of selection.
Sample size	271. The number of individuals in a sample. The sample size determines how well the sample represents the population, not the fraction of the population sampled.
Census	271. A sample that consists of the entire population is called a census.
Population parameter	272. A numerically valued attribute of a model for a population. We rarely expect to know the true value of a population parameter, but we do hope to estimate it from sampled data. For example, the mean income of all employed people in the country is a population parameter.
Statistic, sample statistic	272. Statistics are values calculated for sampled data. Those that correspond to, and thus estimate, a population parameter, are of particular interest. For example, the mean income of all employed people in a representative sample can provide a good estimate of the corresponding population parameter. The term “sample statistic” is sometimes used, usually to parallel the corresponding term “population parameter.”
Representative	273. A sample is said to be representative if the statistics computed from it accurately reflect the corresponding population parameters.
Simple random sample (SRS)	273. A simple random sample of sample size $n$ is a sample in which each set of $n$ elements in the population has an equal chance of selection.
Sampling frame	273. A list of individuals from whom the sample is drawn is called the sampling frame. Individuals who may be in the population of interest, but who are not in the sampling frame, cannot be included in any sample.
Sampling variability	274. The natural tendency of randomly drawn samples to differ, one from another. Sometimes, unfortunately, called <i>sampling error</i> , sampling variability is no error at all, but just the natural result of random sampling.
Stratified random sample	274. A sampling design in which the population is divided into several subpopulations, or <b>strata</b> , and random samples are then drawn from each stratum. If the strata are homogeneous, but are different from each other, a stratified sample may yield more consistent results than an SRS.
Cluster sample	275. A sampling design in which entire groups, or <b>clusters</b> , are chosen at random. Cluster sampling is usually selected as a matter of convenience, practicality, or cost. Each cluster should be representative of the population, so all the clusters should be heterogeneous and similar to each other.
Multistage sample	276. Sampling schemes that combine several sampling methods are called multistage samples. For example, a national polling service may stratify the country by geographical regions, select a random sample of cities from each region, and then interview a cluster of residents in each city.
Systematic sample	277. A sample drawn by selecting individuals systematically from a sampling frame. When there is no relationship between the order of the sampling frame and the variables of interest, a systematic sample can be representative.
Pilot	281. A small trial run of a survey to check whether questions are clear. A pilot study can reduce errors due to ambiguous questions.
Voluntary response bias	282. Bias introduced to a sample when individuals can choose on their own whether to participate in the sample. Samples based on voluntary response are always invalid and cannot be recovered, no matter how large the sample size.

Convenience sample	282. A convenience sample consists of the individuals who are conveniently available. Convenience samples often fail to be representative because every individual in the population is not equally convenient to sample.
Undercoverage	283. A sampling scheme that biases the sample in a way that gives a part of the population less representation than it has in the population suffers from undercoverage.
Nonresponse bias	283. Bias introduced when a large fraction of those sampled fails to respond. Those who do respond are likely to not represent the entire population. Voluntary response bias is a form of nonresponse bias, but nonresponse may occur for other reasons. For example, those who are at work during the day won't respond to a telephone survey conducted only during working hours.
Response bias	284. Anything in a survey design that influences responses falls under the heading of response bias. One typical response bias arises from the wording of questions, which may suggest a favored response. Voters, for example, are more likely to express support of "the president" than support of the particular person holding that office at the moment.

## Skills

### THINK

- ▶ Know the basic concepts and terminology of sampling (see the preceding list).
- ▶ Recognize population parameters in descriptions of populations and samples.
- ▶ Understand the value of randomization as a defense against bias.
- ▶ Understand the value of sampling to estimate population parameters from statistics calculated on representative samples drawn from the population.
- ▶ Understand that the size of the sample (not the fraction of the population) determines the precision of estimates.

### SHOW

- ▶ Know how to draw a simple random sample from a master list of a population, using a computer or a table of random numbers.

### TELL

- ▶ Know what to report about a sample as part of your account of a statistical analysis.
- ▶ Report possible sources of bias in sampling methods. Recognize voluntary response and nonresponse as sources of bias in a sample survey.

## SAMPLING ON THE COMPUTER

Computer-generated pseudorandom numbers are usually good enough for drawing random samples. But there is little reason not to use the truly random values available on the Internet.

Here's a convenient way to draw an SRS of a specified size using a computer-based sampling frame. The sampling frame can be a list of names or of identification numbers arrayed, for example, as a column in a spreadsheet, statistics program, or database:

1. Generate random numbers of enough digits so that each exceeds the size of the sampling frame list by several digits. This makes duplication unlikely.
2. Assign the random numbers arbitrarily to individuals in the sampling frame list. For example, put them in an adjacent column.
3. Sort the list of random numbers, carrying along the sampling frame list.
4. Now the first  $n$  values in the sorted sampling frame column are an SRS of  $n$  values from the entire sampling frame.

## EXERCISES

- Roper.** Through their *Roper Reports Worldwide*, GfK Roper conducts a global consumer survey to help multinational companies understand different consumer attitudes throughout the world. Within 30 countries, the researchers interview 1000 people aged 13–65. Their samples are designed so that they get 500 males and 500 females in each country. ([www.gfkamerica.com](http://www.gfkamerica.com))
  - Are they using a simple random sample? Explain.
  - What kind of design do you think they are using?
- Student Center Survey.** For their class project, a group of Statistics students decide to survey the student body to assess opinions about the proposed new student center. Their sample of 200 contained 50 first-year students, 50 sophomores, 50 juniors, and 50 seniors.
  - Do you think the group was using an SRS? Why?
  - What sampling design do you think they used?
- Emoticons.** The Web site [www.gamefaqs.com](http://www.gamefaqs.com) asked, as their question of the day to which visitors to the site were invited to respond, “Do you ever use emoticons when you type online?” Of the 87,262 respondents, 27% said that they did not use emoticons. ;-(
  - What kind of sample was this?
  - How much confidence would you place in using 27% as an estimate of the fraction of people who use emoticons?
- Drug tests.** Major League Baseball tests players to see whether they are using performance-enhancing drugs. Officials select a team at random, and a drug-testing crew shows up unannounced to test all 40 players on the team. Each testing day can be considered a study of drug use in Major League Baseball.
  - What kind of sample is this?
  - Is that choice appropriate?
- Gallup.** At its Web site ([www.gallup.com](http://www.gallup.com)) the Gallup Poll publishes results of a new survey each day. Scroll down to the end, and you’ll find a statement that includes words such as these:
 

*Results are based on telephone interviews with 1,008 national adults, aged 18 and older, conducted April 2–5, 2007. . . . In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.*

  - For this survey, identify the population of interest.
  - Gallup performs its surveys by phoning numbers generated at random by a computer program. What is the sampling frame?
  - What problems, if any, would you be concerned about in matching the sampling frame with the population?
- Gallup World.** At its Web site ([www.gallupworldpoll.com](http://www.gallupworldpoll.com)) the Gallup World Poll describes their methods. After one report they explained:

*Results are based on face-to-face interviews with randomly selected national samples of approximately 1,000 adults,*

*aged 15 and older, who live permanently in each of the 21 sub-Saharan African nations surveyed. Those countries include Angola (areas where land mines might be expected were excluded), Benin, Botswana, Burkina Faso, Cameroon, Ethiopia, Ghana, Kenya, Madagascar (areas where interviewers had to walk more than 20 kilometers from a road were excluded), Mali, Mozambique, Niger, Nigeria, Senegal, Sierra Leone, South Africa, Tanzania, Togo, Uganda (the area of activity of the Lord’s Resistance Army was excluded from the survey), Zambia, and Zimbabwe. . . . In all countries except Angola, Madagascar, and Uganda, the sample is representative of the entire population.*

- Gallup is interested in sub-Saharan Africa. What kind of survey design are they using?
- Some of the countries surveyed have large populations. (Nigeria is estimated to have about 130 million people.) Some are quite small. (Togo’s population is estimated at 5.4 million.) Nonetheless, Gallup sampled 1000 adults in each country. How does this affect the precision of its estimates for these countries?

**7–14. What did they do?** For the following reports about statistical studies, identify the following items (if possible). If you can’t tell, then say so—this often happens when we read about a survey.

- The population
  - The population parameter of interest
  - The sampling frame
  - The sample
  - The sampling method, including whether or not randomization was employed
  - Any potential sources of bias you can detect and any problems you see in generalizing to the population of interest
- Consumers Union asked all subscribers whether they had used alternative medical treatments and, if so, whether they had benefited from them. For almost all of the treatments, approximately 20% of those responding reported cures or substantial improvement in their condition.
  - A question posted on the Lycos Web site on 18 June 2000 asked visitors to the site to say whether they thought that marijuana should be legally available for medicinal purposes. ([www.lycos.com](http://www.lycos.com))
  - Researchers waited outside a bar they had randomly selected from a list of such establishments. They stopped every 10th person who came out of the bar and asked whether he or she thought drinking and driving was a serious problem.
  - Hoping to learn what issues may resonate with voters in the coming election, the campaign director for a mayoral candidate selects one block from each of the city’s election districts. Staff members go there and interview all the residents they can find.

11. The Environmental Protection Agency took soil samples at 16 locations near a former industrial waste dump and checked each for evidence of toxic chemicals. They found no elevated levels of any harmful substances.
12. State police set up a roadblock to estimate the percentage of cars with up-to-date registration, insurance, and safety inspection stickers. They usually find problems with about 10% of the cars they stop.
13. A company packaging snack foods maintains quality control by randomly selecting 10 cases from each day's production and weighing the bags. Then they open one bag from each case and inspect the contents.
14. Dairy inspectors visit farms unannounced and take samples of the milk to test for contamination. If the milk is found to contain dirt, antibiotics, or other foreign matter, the milk will be destroyed and the farm reinspected until purity is restored.
15. **Mistaken poll.** A local TV station conducted a "PulsePoll" about the upcoming mayoral election. Evening news viewers were invited to phone in their votes, with the results to be announced on the late-night news. Based on the phone calls, the station predicted that Amabo would win the election with 52% of the vote. They were wrong: Amabo lost, getting only 46% of the vote. Do you think the station's faulty prediction is more likely to be a result of bias or sampling error? Explain.
16. **Another mistaken poll.** Prior to the mayoral election discussed in Exercise 15, the newspaper also conducted a poll. The paper surveyed a random sample of registered voters stratified by political party, age, sex, and area of residence. This poll predicted that Amabo would win the election with 52% of the vote. The newspaper was wrong: Amabo lost, getting only 46% of the vote. Do you think the newspaper's faulty prediction is more likely to be a result of bias or sampling error? Explain.
17. **Parent opinion, part 1.** In a large city school system with 20 elementary schools, the school board is considering the adoption of a new policy that would require elementary students to pass a test in order to be promoted to the next grade. The PTA wants to find out whether parents agree with this plan. Listed below are some of the ideas proposed for gathering data. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
  - a) Put a big ad in the newspaper asking people to log their opinions on the PTA Web site.
  - b) Randomly select one of the elementary schools and contact every parent by phone.
  - c) Send a survey home with every student, and ask parents to fill it out and return it the next day.
  - d) Randomly select 20 parents from each elementary school. Send them a survey, and follow up with a phone call if they do not return the survey within a week.
18. **Parent opinion, part 2.** Let's revisit the school system described in Exercise 17. Four new sampling strategies have been proposed to help the PTA determine whether parents favor requiring elementary students to pass a test in order to be promoted to the next grade. For each, indicate what kind of sampling strategy is involved and what (if any) biases might result.
  - a) Run a poll on the local TV news, asking people to dial one of two phone numbers to indicate whether they favor or oppose the plan.
  - b) Hold a PTA meeting at each of the 20 elementary schools, and tally the opinions expressed by those who attend the meetings.
  - c) Randomly select one class at each elementary school and contact each of those parents.
  - d) Go through the district's enrollment records, selecting every 40th parent. PTA volunteers will go to those homes to interview the people chosen.
19. **Churches.** For your political science class, you'd like to take a survey from a sample of all the Catholic Church members in your city. A list of churches shows 17 Catholic churches within the city limits. Rather than try to obtain a list of all members of all these churches, you decide to pick 3 churches at random. For those churches, you'll ask to get a list of all current members and contact 100 members at random.
  - a) What kind of design have you used?
  - b) What could go wrong with your design?
20. **Playground.** Some people have been complaining that the children's playground at a municipal park is too small and is in need of repair. Managers of the park decide to survey city residents to see if they believe the playground should be rebuilt. They hand out questionnaires to parents who bring children to the park. Describe possible biases in this sample.
21. **Roller coasters.** An amusement park has opened a new roller coaster. It is so popular that people are waiting for up to 3 hours for a 2-minute ride. Concerned about how patrons (who paid a large amount to enter the park and ride on the rides) feel about this, they survey every 10th person on the line for the roller coaster, starting from a randomly selected individual.
  - a) What kind of sample is this?
  - b) What is the sampling frame?
  - c) Is it likely to be representative?
22. **Playground, act two.** The survey described in Exercise 20 asked,
 

*Many people believe this playground is too small and in need of repair. Do you think the playground should be repaired and expanded even if that means raising the entrance fee to the park?*

 Describe two ways this question may lead to response bias.
23. **Wording the survey.** Two members of the PTA committee in Exercises 17 and 18 have proposed different questions to ask in seeking parents' opinions.
 

**Question 1:** *Should elementary school-age children have to pass high-stakes tests in order to remain with their classmates?*

**Question 2:** *Should schools and students be held accountable for meeting yearly learning goals by testing students before they advance to the next grade?*

  - a) Do you think responses to these two questions might differ? How? What kind of bias is this?
  - b) Propose a question with more neutral wording that might better assess parental opinion.

24. **Banning ephedra.** An online poll at a Web site asked:

*A nationwide ban of the diet supplement ephedra went into effect recently. The herbal stimulant has been linked to 155 deaths and many more heart attacks and strokes. Ephedra manufacturer NVE Pharmaceuticals, claiming that the FDA lacked proof that ephedra is dangerous if used as directed, was denied a temporary restraining order on the ban yesterday by a federal judge. Do you think that ephedra should continue to be banned nationwide?*

65% of 17,303 respondents said “yes.” Comment on each of the following statements about this poll:

- With a sample size that large, we can be pretty certain we know the true proportion of Americans who think ephedra should be banned.
  - The wording of the question is clearly very biased.
  - The sampling frame is all Internet users.
  - Results of this voluntary response survey can't be reliably generalized to any population of interest.
25. **Survey questions.** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.
- Should companies that pollute the environment be compelled to pay the costs of cleanup?
  - Given that 18-year-olds are old enough to vote and to serve in the military, is it fair to set the drinking age at 21?
26. **More survey questions.** Examine each of the following questions for possible bias. If you think the question is biased, indicate how and propose a better question.
- Do you think high school students should be required to wear uniforms?
  - Given humanity's great tradition of exploration, do you favor continued funding for space flights?
27. **Phone surveys.** Anytime we conduct a survey, we must take care to avoid undercoverage. Suppose we plan to select 500 names from the city phone book, call their homes between noon and 4 p.m., and interview whoever answers, anticipating contacts with at least 200 people.
- Why is it difficult to use a simple random sample here?
  - Describe a more convenient, but still random, sampling strategy.
  - What kinds of households are likely to be included in the eventual sample of opinion? Excluded?
  - Suppose, instead, that we continue calling each number, perhaps in the morning or evening, until an adult is contacted and interviewed. How does this improve the sampling design?
  - Random-digit dialing machines can generate the phone calls for us. How would this improve our design? Is anyone still excluded?
28. **Cell phone survey.** What about drawing a random sample only from cell phone exchanges? Discuss the advantages and disadvantages of such a sampling method compared with surveying randomly generated telephone numbers from non-cell phone exchanges. Do you think these advantages and disadvantages have changed over time? How do you expect they'll change in the future?

29. **Arm length.** How long is your arm compared with your hand size? Put your right thumb at your left shoulder bone, stretch your hand open wide, and extend your hand down your arm. Put your thumb at the place where your little finger is, and extend down the arm again. Repeat this a third time. Now your little finger will probably have reached the back of your left hand. If the fourth hand width goes past the end of your middle finger, turn your hand sideways and count finger widths to get there.
- How many hand and finger widths is your arm?
  - Suppose you repeat your measurement 10 times and average your results. What parameter would this average estimate? What is the population?
  - Suppose you now collect arm lengths measured in this way from 9 friends and average these 10 measurements. What is the population now? What parameter would this average estimate?
  - Do you think these 10 arm lengths are likely to be representative of the population of arm lengths in your community? In the country? Why or why not?
30. **Fuel economy.** Occasionally, when I fill my car with gas, I figure out how many miles per gallon my car got. I wrote down those results after 6 fill-ups in the past few months. Overall, it appears my car gets 28.8 miles per gallon.
- What statistic have I calculated?
  - What is the parameter I'm trying to estimate?
  - How might my results be biased?
  - When the Environmental Protection Agency (EPA) checks a car like mine to predict its fuel economy, what parameter is it trying to estimate?
31. **Accounting.** Between quarterly audits, a company likes to check on its accounting procedures to address any problems before they become serious. The accounting staff processes payments on about 120 orders each day. The next day, the supervisor rechecks 10 of the transactions to be sure they were processed properly.
- Propose a sampling strategy for the supervisor.
  - How would you modify that strategy if the company makes both wholesale and retail sales, requiring different bookkeeping procedures?
32. **Happy workers?** A manufacturing company employs 14 project managers, 48 foremen, and 377 laborers. In an effort to keep informed about any possible sources of employee discontent, management wants to conduct job satisfaction interviews with a sample of employees every month.
- Do you see any potential danger in the company's plan? Explain.
  - Propose a sampling strategy that uses a simple random sample.
  - Why do you think a simple random sample might not provide the representative opinion the company seeks?
  - Propose a better sampling strategy.
  - Listed below are the last names of the project managers. Use random numbers to select two people to be interviewed. Explain your method carefully.

Barrett	Bowman	Chen
DeLara	DeRoos	Grigorov
Maceli	Mulvaney	Pagliariulo
Rosica	Smithson	Tadros
Williams	Yamamoto	



33. **Quality control.** Sammy's Salsa, a small local company, produces 20 cases of salsa a day. Each case contains 12 jars and is imprinted with a code indicating the date and batch number. To help maintain consistency, at the end of each day, Sammy selects three jars of salsa, weighs the contents, and tastes the product. Help Sammy select the sample jars. Today's cases are coded 07N61 through 07N80.
- Carefully explain your sampling strategy.
  - Show how to use random numbers to pick 3 jars.
  - Did you use a simple random sample? Explain.
34. **A fish story.** Concerned about reports of discolored scales on fish caught downstream from a newly sited chemical plant, scientists set up a field station in a shoreline public park. For one week they asked fishermen there to bring any fish they caught to the field station for a brief inspection. At the end of the week, the scientists said that 18% of the 234 fish that were submitted for inspection displayed the discoloration. From this information, can the researchers estimate what proportion of fish in the river have discolored scales? Explain.
35. **Sampling methods.** Consider each of these situations. Do you think the proposed sampling method is appropriate? Explain.
- We want to know what percentage of local doctors accept Medicaid patients. We call the offices of 50 doctors randomly selected from local Yellow Page listings.
  - We want to know what percentage of local businesses anticipate hiring additional employees in the upcoming month. We randomly select a page in the Yellow Pages and call every business listed there.
36. **More sampling methods.** Consider each of these situations. Do you think the proposed sampling method is appropriate? Explain.
- We want to know if there is neighborhood support to turn a vacant lot into a playground. We spend a Saturday afternoon going door-to-door in the neighborhood, asking people to sign a petition.
  - We want to know if students at our college are satisfied with the selection of food available on campus. We go to the largest cafeteria and interview every 10th person in line.



## JUST CHECKING Answers

- It can be hard to reach all members of a population, and it can take so long that circumstances change, affecting the responses. A well-designed sample is often a better choice.
  - This sample is probably biased—students who didn't like the food at the cafeteria might not choose to eat there.
  - No, only the sample size matters, not the fraction of the overall population.
  - Students who frequent this Web site might be more enthusiastic about Statistics than the overall population of Statistics students. A large sample cannot compensate for bias.
  - It's the population "parameter." "Statistics" describe samples.
- systematic
  - stratified
  - simple
  - cluster

# Experiments and Observational Studies



Who gets good grades? And, more importantly, why? Is there something schools and parents could do to help weaker students improve their grades? Some people think they have an answer: music! No, not your iPod, but an instrument. In a study conducted at Mission Viejo High School, in California, researchers compared the scholastic performance of music students with that of non-music students. Guess what? The music students had a much higher overall grade point average than the non-music students, 3.59 to 2.91. Not only that: A whopping 16% of the music students had all A's compared with only 5% of the non-music students.

As a result of this study and others, many parent groups and educators pressed for expanded music programs in the nation's schools. They argued that the work ethic, discipline, and feeling of accomplishment fostered by learning to play an instrument also enhance a person's ability to succeed in school. They thought that involving more students in music would raise academic performance. What do you think? Does this study provide solid evidence? Or are there other possible explanations for the difference in grades? Is there any way to really prove such a conjecture?

## Observational Studies

This research tried to show an association between music education and grades. But it wasn't a survey. Nor did it assign students to get music education. Instead, it simply observed students "in the wild," recording the choices they made and the outcome. Such studies are called **observational studies**. In observational studies, researchers don't *assign* choices; they simply observe them. In addition, this was a **retrospective study**, because researchers first identified subjects who studied music and then collected data on their past grades.

What's wrong with concluding that music education causes good grades? One high school during one academic year may not be representative of the

whole United States. That's true, but the real problem is that the claim that music study *caused* higher grades depends on there being *no other differences* between the groups that could account for the differences in grades, and studying music was not the *only* difference between the two groups of students.

We can think of lots of lurking variables that might cause the groups to perform differently. Students who study music may have better work habits to start with, and this makes them successful in both music and course work. Music students may have more parental support (someone had to pay for all those lessons), and that support may have enhanced their academic performance, too. Maybe they came from wealthier homes and had other advantages. Or it could be that smarter kids just like to play musical instruments.

For rare illnesses, it's not practical to draw a large enough sample to see many ill respondents, so the only option remaining is to develop retrospective data. For example, researchers can interview those who have become ill. The likely causes of both legionnaires' disease and HIV were initially identified from such retrospective studies of the small populations who were initially infected. But to confirm the causes, researchers needed laboratory-based experiments.

Observational studies are valuable for discovering trends and possible relationships. They are used widely in public health and marketing. Observational studies that try to discover variables related to rare outcomes, such as specific diseases, are often retrospective. They first identify people with the disease and then look into their history and heritage in search of things that may be related to their condition. But retrospective studies have a restricted view of the world because they are usually restricted to a small part of the entire population. **And because retrospective records are based on historical data, they can have errors.** (Do you recall *exactly* what you ate even yesterday? How about last Wednesday?)

A somewhat better approach is to observe individuals over time, recording the variables of interest and ultimately seeing how things turn out. For example, we might start by selecting young students who have not begun music lessons. We could then track their academic performance over several years, comparing those who later choose to study music with those who do not. **Identifying subjects in advance and collecting data as events unfold would make this a prospective study.**

Although an observational study may identify important variables related to the outcome we are interested in, there is no guarantee that we have found the right or the most important related variables. Students who choose to study an instrument might still differ from the others in some important way that we failed to observe. It may be this difference—whether we know what it is or not—rather than music itself that leads to better grades. It's just not possible for observational studies, whether prospective or retrospective, to demonstrate a causal relationship.

## FOR EXAMPLE

### Designing an observational study

In early 2007, a larger-than-usual number of cats and dogs developed kidney failure; many died. Initially, researchers didn't know why, so they used an observational study to investigate.

**Question:** Suppose you were called on to plan a study seeking the cause of this problem. Would your design be retrospective or prospective? Explain why.

I would use a retrospective observational study. Even though the incidence of disease was higher than usual, it was still rare. Surveying all pets would have been impractical. Instead, it makes sense to locate some who were sick and ask about their diets, exposure to toxins, and other possible causes.



## Randomized, Comparative Experiments



Experimental design was advanced in the 19th century by work in psychophysics by Gustav Fechner (1801–1887), the founder of experimental psychology. Fechner designed ingenious experiments that exhibited many of the features of modern designed experiments. Fechner was careful to control for the effects of factors that might affect his results. For example, in his 1860 book *Elemente der Psychophysik* he cautioned readers to group experiment trials together to minimize the possible effects of time of day and fatigue.

### An Experiment:

Manipulates the factor levels to create treatments.  
Randomly assigns subjects to these treatment levels.  
Compares the responses of the subject groups across treatment levels.

*“He that leaves nothing to chance will do few things ill, but he will do very few things.”*

—Lord Halifax  
(1633–1695)

Is it *ever* possible to get convincing evidence of a cause-and-effect relationship? Well, yes it is, but we would have to take a different approach. We could take a group of third graders, randomly assign half to take music lessons, and forbid the other half to do so. Then we could compare their grades several years later. **This kind of study design is called an experiment.**

An experiment requires a **random assignment** of subjects to treatments. Only an experiment can justify a claim like “Music lessons cause higher grades.” Questions such as “Does taking vitamin C reduce the chance of getting a cold?” and “Does working with computers improve performance in Statistics class?” and “Is this drug a safe and effective treatment for that disease?” require a designed experiment to establish cause and effect.

Experiments study the relationship between two or more variables. An experimenter must identify at least one explanatory variable, called a **factor**, to manipulate and at least one **response variable** to measure. What distinguishes an experiment from other types of investigation is that the experimenter actively and deliberately manipulates the factors to control the details of the possible treatments, and assigns the subjects to those treatments *at random*. The experimenter then observes the response variable and *compares* responses for different groups of subjects who have been treated differently. For example, we might design an experiment to see whether the amount of sleep and exercise you get affects your performance.

The individuals on whom or which we experiment are known by a variety of terms. Humans who are experimented on are commonly called **subjects** or **participants**. Other individuals (rats, days, petri dishes of bacteria) are commonly referred to by the more generic term **experimental unit**. When we recruit subjects for our sleep deprivation experiment by advertising in Statistics class, we’ll probably have better luck if we invite them to be participants than if we advertise that we need experimental units.

The specific values that the experimenter chooses for a factor are called the **levels** of the factor. We might assign our participants to sleep for 4, 6, or 8 hours. Often there are several factors at a variety of levels. (Our subjects will also be assigned to a treadmill for 0 or 30 minutes.) The combination of specific levels from all the factors that an experimental unit receives is known as its **treatment**. (Our subjects could have any one of six different treatments—three sleep levels, each at two exercise levels.)

How should we assign our participants to these treatments? Some students prefer 4 hours of sleep, while others need 8. Some exercise regularly; others are couch potatoes. Should we let the students choose the treatments they’d prefer? No. That would not be a good idea. To have any hope of drawing a fair conclusion, we must assign our participants to their treatments *at random*.

It may be obvious to you that we shouldn’t let the students choose the treatment they’d prefer, but the need for random assignment is a lesson that was once hard for some to accept. For example, physicians might naturally prefer to assign patients to the therapy that they think best rather than have a random element such as a coin flip determine the treatment. But we’ve known for more than a century that for the results of an experiment to be valid, we must use deliberate randomization.

**The Women’s Health Initiative** is a major 15-year research program funded by the National Institutes of Health to address the most common causes of death, disability, and poor quality of life in older women. It consists of both an observational study with more than 93,000 participants and several randomized comparative experiments. The goals of this study include

- ▶ giving reliable estimates of the extent to which known risk factors predict heart disease, cancers, and fractures;

No drug can be sold in the United States without first showing, in a suitably designed experiment approved by the Food and Drug Administration (FDA), that it's safe and effective. The small print on the booklet that comes with many prescription drugs usually describes the outcomes of that experiment.

- ▶ identifying “new” risk factors for these and other diseases in women;
- ▶ comparing risk factors, presence of disease at the start of the study, and new occurrences of disease during the study across all study components; and
- ▶ creating a future resource to identify biological indicators of disease, especially substances and factors found in blood.

That is, the study seeks to identify possible risk factors and assess how serious they might be. It seeks to build up data that might be checked retrospectively as the women in the study continue to be followed. There would be no way to find out these things with an experiment because the task includes identifying new risk factors. If we don't know those risk factors, we could never control them as factors in an experiment.

By contrast, one of the clinical trials (randomized experiments) that received much press attention randomly assigned postmenopausal women to take either hormone replacement therapy or an inactive pill. The results published in 2002 and 2004 concluded that hormone replacement with estrogen carried increased risks of stroke.

## FOR EXAMPLE

### Determining the treatments and response variable

**Recap:** In 2007, deaths of a large number of pet dogs and cats were ultimately traced to contamination of some brands of pet food. The manufacturer now claims that the food is safe, but before it can be released, it must be tested.

**Question:** In an experiment to test whether the food is now safe for dogs to eat,<sup>1</sup> what would be the treatments and what would be the response variable?

*The treatments would be ordinary-size portions of two dog foods: the new one from the company (the test food) and one that I was certain was safe (perhaps prepared in my kitchen or laboratory). The response would be a veterinarian's assessment of the health of the test animals.*

# The Four Principles of Experimental Design

AS

**Video: An Industrial Experiment.** Manufacturers often use designed experiments to help them perfect new products. Watch this video about one such experiment.

1. **Control.** We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. For human subjects, we try to treat them alike. However, there is always a question of degree and practicality. Controlling extraneous sources of variation reduces the variability of the responses, making it easier to detect differences among the treatment groups.

Making generalizations from the experiment to other levels of the controlled factor can be risky. For example, suppose we test two laundry detergents and carefully control the water temperature at 180°F. This would reduce the variation in our results due to water temperature, but what could we say about the detergents' performance in cold water? Not much. It would be hard to justify extrapolating the results to other temperatures.

Although we control both experimental factors and other sources of variation, we think of them very differently. We control a factor by assigning subjects to different factor levels because we want to see how the response will change at those different levels. We control other sources of variation to *prevent* them from changing and affecting the response variable.

<sup>1</sup> It may disturb you (as it does us) to think of deliberately putting dogs at risk in this experiment, but in fact that is what is done. The risk is borne by a small number of dogs so that the far larger population of dogs can be kept safe.

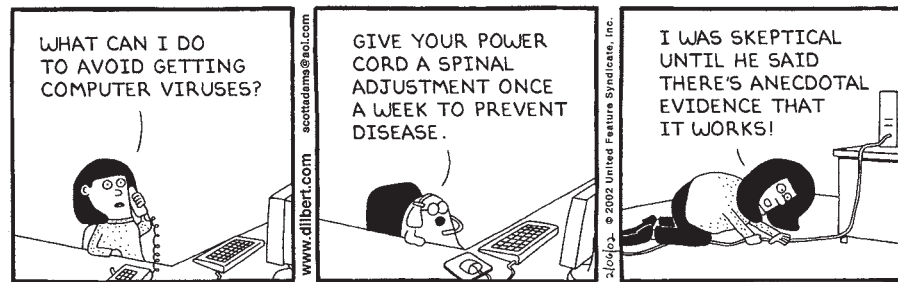


The deep insight that experiments should use random assignment is quite an old one. It can be attributed to the American philosopher and scientist C. S. Peirce in his experiments with J. Jastrow, published in 1885.

**AS** **Activity: The Three Rules of Experimental Design.** Watch an animated discussion of three rules of design.

**AS** **Activity: Perform an Experiment.** How well can you read pie charts and bar charts? Find out as you serve as the subject in your own experiment.

- 2. Randomize.** As in sample surveys, **randomization** allows us to equalize the effects of unknown or uncontrollable sources of variation. It does not eliminate the effects of these sources, but it should spread them out across the treatment levels so that we can see past them. If experimental units were not assigned to treatments at random, we would not be able to use the powerful methods of Statistics to draw conclusions from an experiment. Assigning subjects to treatments at random reduces bias due to uncontrolled sources of variation. Randomization protects us even from effects we didn't know about. There's an adage that says "control what you can, and randomize the rest."
- 3. Replicate.** Two kinds of replication show up in comparative experiments. First, we should apply each treatment to a number of subjects. Only with such replication can we estimate the variability of responses. If we have not assessed the variation, the experiment is not complete. The outcome of an experiment on a single subject is an anecdote, not data.



A second kind of replication shows up when the experimental units are not a representative sample from the population of interest. We may believe that what is true of the students in Psych 101 who volunteered for the sleep experiment is true of all humans, but we'll feel more confident if our results for the experiment are *replicated* in another part of the country, with people of different ages, and at different times of the year. **Replication of an entire experiment with the controlled sources of variation at different levels is an essential step in science.**

- 4. Block.** The ability of randomizing to equalize variation across treatment groups works best in the long run. For example, if we're allocating players to two 6-player soccer teams from a pool of 12 children, we might do so at random to equalize the talent. But what if there were two 12-year-olds and ten 6-year-olds in the group? Randomizing may place both 12-year-olds on the same team. In the long run, if we did this over and over, it would all equalize. But wouldn't it be better to assign one 12-year-old to each group (at random) and five 6-year-olds to each team (at random)? By doing this, we would improve fairness in the short run. This approach makes the division more fair by recognizing the variation in *age* and allocating the players at random *within* each age level. When we do this, we call the variable *age* a **blocking variable**. The levels of *age* are called blocks.

Sometimes, attributes of the experimental units that we are not studying and that we can't control may nevertheless affect the outcomes of an experiment. If we group similar individuals together and then randomize within each of these **blocks**, we can remove much of the variability due to the difference among the blocks. Blocking is an important compromise between randomization and control. However, unlike the first three principles, blocking is not *required* in an experimental design.

## FOR EXAMPLE

## Control, randomize, and replicate

**Recap:** We're planning an experiment to see whether the new pet food is safe for dogs to eat. We'll feed some animals the new food and others a food known to be safe, comparing their health after a period of time.

**Questions:** In this experiment, how will you implement the principles of control, randomization, and replication?

I'd control the portion sizes eaten by the dogs. To reduce possible variability from factors other than the food, I'd standardize other aspects of their environments—housing the dogs in similar pens and ensuring that each got the same amount of water, exercise, play, and sleep time, for example. I might restrict the experiment to a single breed of dog and to adult dogs to further minimize variation.

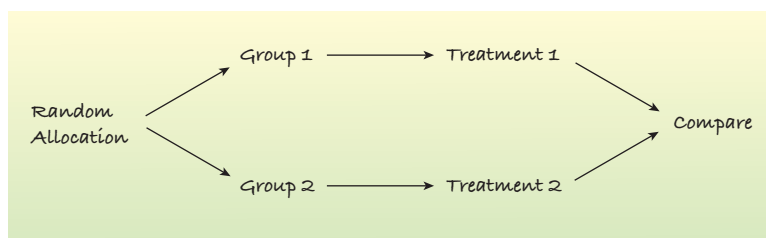
To equalize traits, pre-existing conditions, and other unknown influences, I would assign dogs to the two feed treatments randomly.

I would replicate by assigning more than one dog to each treatment to allow for variability among individual dogs. If I had the time and funding, I might replicate the entire experiment using, for example, a different breed of dog.



## Diagrams

An experiment is carried out over time with specific actions occurring in a specified order. A diagram of the procedure can help in thinking about experiments.<sup>2</sup>



The diagram emphasizes the random allocation of subjects to treatment groups, the separate treatments applied to these groups, and the ultimate comparison of results. It's best to specify the responses that will be compared. A good way to start comparing results for the treatment groups is with boxplots.

## STEP-BY-STEP EXAMPLE

## Designing an Experiment



An ad for OptiGro plant fertilizer claims that with this product you will grow “juicier, tastier” tomatoes. You'd like to test this claim, and wonder whether you might be able to get by with half the specified dose. How can you set up an experiment to check out the claim?

Of course, you'll have to get some tomatoes, try growing some plants with the product and some without, and see what happens. But you'll need a clearer plan than that. How should you design your experiment?

<sup>2</sup> Diagrams of this sort were introduced by David Moore in his textbooks and are widely used.

A completely randomized experiment is the ideal simple design, just as a *simple random sample* is the ideal simple sample—and for many of the same reasons.

Let's work through the design, step by step. We'll design the simplest kind of experiment, a **completely randomized experiment in one factor**. Since this is a *design* for an experiment, most of the steps are part of the *Think* stage. The statements in the right column are the kinds of things you would need to say in *proposing* an experiment. You'd need to include them in the "methods" section of a report once the experiment is run.

**Question:** How would you design an experiment to test OptiGro fertilizer?



**Plan** State what you want to know.

I want to know whether tomato plants grown with OptiGro yield juicier, tastier tomatoes than plants raised in otherwise similar circumstances but without the fertilizer.

**Response** Specify the response variable.

I'll evaluate the juiciness and taste of the tomatoes by asking a panel of judges to rate them on a scale from 1 to 7 in juiciness and in taste.

**Treatments** Specify the factor levels and the treatments.

The factor is fertilizer, specifically OptiGro fertilizer. I'll grow tomatoes at three different factor levels: some with no fertilizer, some with half the specified amount of OptiGro, and some with the full dose of OptiGro. These are the three treatments.

**Experimental Units** Specify the experimental units.

I'll obtain 24 tomato plants of the same variety from a local garden store.

**Experimental Design** Observe the principles of design:

**Control** any sources of variability you know of and can control.

I'll locate the farm plots near each other so that the plants get similar amounts of sun and rain and experience similar temperatures. I will weed the plots equally and otherwise treat the plants alike.

**Replicate** results by placing more than one plant in each treatment group.

I'll use 8 plants in each treatment group.

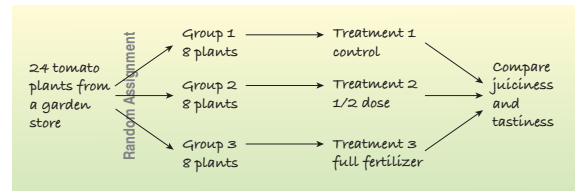
**Randomly assign** experimental units to treatments, to equalize the effects of unknown or uncontrollable sources of variation.

To randomly divide the plants into three groups, first I'll label the plants with numbers 00–23. I'll look at pairs of digits across a random number table. The first 8 plants identified (ignoring numbers 24–99 and any repeats) will go in Group 1, the next 8 in Group 2, and the remaining plants in Group 3.

Describe how the randomization will be accomplished.



**Make a Picture** A diagram of your design can help you think about it clearly.



Specify any other experiment details. You must give enough details so that another experimenter could exactly replicate your experiment. It's generally better to include details that might seem irrelevant than to leave out matters that could turn out to make a difference.

I will grow the plants until the tomatoes are mature, as judged by reaching a standard color.

I'll harvest the tomatoes when ripe and store them for evaluation.

Specify how to measure the response.

I'll set up a numerical scale of juiciness and one of tastiness for the taste testers. Several people will taste slices of tomato and rate them.

SHOW

Once you collect the data, you'll need to display them and compare the results for the three treatment groups.

I will display the results with side-by-side boxplots to compare the three treatment groups.

I will compare the means of the groups.

TELL

To answer the initial question, we ask whether the differences we observe in the means of the three groups are meaningful.

If the differences in taste and juiciness among the groups are greater than I would expect by knowing the usual variation among tomatoes, I may be able to conclude that these differences can be attributed to treatment with the fertilizer.

Because this is a randomized experiment, we can attribute significant differences to the treatments. To do this properly, we'll need methods from what is called "statistical inference," the subject of the rest of this book.

## Does the Difference Make a Difference?

If the differences among the treatment groups are big enough, we'll attribute the differences to the treatments, but how can we decide whether the differences are big enough?

Would we expect the group means to be identical? Not really. Even if the treatment made no difference whatever, there would still be some variation. We assigned the tomato plants to treatments at random. But a different random assignment would have led to different results. Even a repeat of the *same* treatment on a different randomly assigned set of plants would lead to a different mean. The real question is whether the differences we observed are about as big as we might get just from the randomization alone, or whether they're bigger than that. If we decide that they're bigger, we'll attribute the differences to the treatments. In that case we say the differences are **statistically significant**.

**A S**

### Activity: Graph the Data.

Do you think there's a significant difference in your perception of pie charts and bar charts? Explore the data from your plot perception experiment.

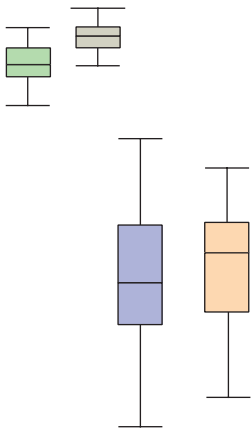
How will we decide if something is different enough to be considered statistically significant? Later chapters will offer methods to help answer that question, but to get some intuition, think about deciding whether a coin is fair. If we flip a fair coin 100 times, we expect, *on average*, to get 50 heads. Suppose we get 54 heads out of 100. That doesn't seem very surprising. It's well within the bounds of ordinary random fluctuations. What if we'd seen 94 heads? That's clearly outside the bounds. We'd be pretty sure that the coin flips were not random. But what about 74 heads? Is that far enough from 50% to arouse our suspicions? That's the sort of question we need to ask of our experiment results.

In Statistics terminology, 94 heads would be a statistically significant difference from 50, and 54 heads would not. Whether 74 is *statistically significant* or not would depend on the chance of getting 74 heads in 100 flips of a fair coin and on our tolerance for believing that rare events can happen to us.

Back at the tomato stand, we ask whether the differences we see among the treatment groups are the kind of differences we'd expect from randomization. A good way to get a feeling for that is to look at how much our results vary among plants that get the *same* treatment. Boxplots of our results by treatment group can give us a general idea.

For example, Figure 13.1 shows two pairs of boxplots whose centers differ by exactly the same amount. In the upper set, that difference appears to be larger than we'd expect just by chance. Why? Because the variation is quite small *within* treatment groups, so the larger difference *between* the groups is unlikely to be just from the randomization. In the bottom pair, that same difference between the centers looks less impressive. There the variation *within* each group swamps the difference *between* the two medians. We'd say the difference is statistically significant in the upper pair and not statistically significant in the lower pair.

In later chapters we'll see statistical tests that quantify this intuition. For now, the important point is that a difference is statistically significant if we don't believe that it's likely to have occurred only by chance.



**FIGURE 13.1**

The boxplots in both pairs have centers the same distance apart, but when the spreads are large, the observed difference may be just from random fluctuation.



## JUST CHECKING

- At one time, a method called “gastric freezing” was used to treat people with peptic ulcers. An inflatable bladder was inserted down the esophagus and into the stomach, and then a cold liquid was pumped into the bladder. Now you can find the following notice on the Internet site of a major insurance company:

[Our company] does not cover gastric freezing (intra-gastric hypothermia) for chronic peptic ulcer disease. . . .

Gastric freezing for chronic peptic ulcer disease is a non-surgical treatment which was popular about 20 years ago but now is seldom performed. It has been abandoned due to a high complication rate, only temporary improvement experienced by patients, and a lack of effectiveness when tested by double-blind, controlled clinical trials.

What did that “controlled clinical trial” (experiment) probably look like? (Don't worry about “double-blind”; we'll get to that soon.)

- |                                                   |                                                                               |
|---------------------------------------------------|-------------------------------------------------------------------------------|
| <b>a)</b> What was the factor in this experiment? | <b>d)</b> How did researchers decide which subjects received which treatment? |
| <b>b)</b> What was the response variable?         | <b>e)</b> Were the results statistically significant?                         |
| <b>c)</b> What were the treatments?               |                                                                               |

## Experiments and Samples

Both experiments and sample surveys use randomization to get unbiased data. But they do so in different ways and for different purposes. **Sample surveys try to estimate population parameters**, so the sample needs to be as representative of the population as possible. By contrast, **experiments try to assess the effects of treatments**. Experimental units are not always drawn randomly from the population. For example, a medical experiment may deal only with local patients who

have the disease under study. The randomization is in the assignment of their therapy. We want a sample to exhibit the diversity and variability of the population, but for an experiment the more homogeneous the subjects the more easily we'll spot differences in the effects of the treatments.



Experiments are rarely performed on random samples from a population. Don't describe the subjects in an experiment as a random sample unless they really are. More likely, the randomization was in assigning subjects to treatments.

Unless the experimental units are chosen from the population at random, you should be cautious about generalizing experiment results to larger populations until the experiment has been repeated under different circumstances. Results become more persuasive if they remain the same in completely different settings, such as in a different season, in a different country, or for a different species, to name a few.

Even without choosing experimental units from a population at random, experiments can draw stronger conclusions than surveys. By looking only at the differences across treatment groups, experiments cancel out many sources of bias. For example, the entire pool of subjects may be biased and not representative of the population. (College students may need more sleep, on average, than the general population.) When we assign subjects randomly to treatment groups, all the groups are still biased, but *in the same way*. When we consider the differences in their responses, these biases cancel out, allowing us to see the *differences* due to treatment effects more clearly.

## Control Treatments

**A S**

**Activity: Control Groups in Experiments.** Is a control group really necessary?

Suppose you wanted to test a \$300 piece of software designed to shorten download times. You could just try it on several files and record the download times, but you probably want to *compare* the speed with what would happen *without* the software installed. Such a baseline measurement is called a **control treatment**, and the experimental units to whom it is applied are called a **control group**.

This is a use of the word “control” in an entirely different context. Previously, we controlled extraneous sources of variation by keeping them constant. Here, we use a control treatment as another *level* of the factor in order to compare the treatment results to a situation in which “nothing happens.” That’s what we did in the tomato experiment when we used no fertilizer on the 8 tomatoes in Group 1.

## Blinding

Humans are notoriously susceptible to errors in judgment.<sup>3</sup> All of us. When we know what treatment was assigned, it’s difficult not to let that knowledge influence our assessment of the response, even when we try to be careful.

Suppose you were trying to advise your school on which brand of cola to stock in the school’s vending machines. You set up an experiment to see which of the three competing brands students prefer (or whether they can tell the difference at all). But people have brand loyalties. You probably prefer one brand already. So if you knew which brand you were tasting, it might influence your rating. To avoid this problem, it would be better to disguise the brands as much as possible. This strategy is called **blinding** the participants to the treatment.<sup>4</sup>

But it isn’t just the subjects who should be blind. Experimenters themselves often subconsciously behave in ways that favor what they believe. Even technicians may treat plants or test animals differently if, for example, they expect them to die. An animal that starts doing a little better than others by showing an increased appetite may get fed a bit more than the experimental protocol specifies.

<sup>3</sup> For example, here we are in Chapter 13 and you’re still reading the footnotes.

<sup>4</sup> C. S. Peirce, in the same 1885 work in which he introduced randomization, also recommended blinding.

### Blinding by Misleading

Social science experiments can sometimes blind subjects by misleading them about the purpose of a study. One of the authors participated as an undergraduate volunteer in a (now infamous) psychology experiment using such a blinding method. The subjects were told that the experiment was about three-dimensional spatial perception and were assigned to draw a model of a horse. While they were busy drawing, a loud noise and then groaning were heard coming from the room next door. The *real* purpose of the experiment was to see how people reacted to the apparent disaster. The experimenters wanted to see whether the social pressure of being in groups made people react to the disaster differently. Subjects had been randomly assigned to draw either in groups or alone; that was the treatment. The experimenter had no interest in how well the subjects could draw the horse, but the subjects were blinded to the treatment because they were misled.

People are so good at picking up subtle cues about treatments that the best (in fact, the *only*) defense against such biases in experiments on human subjects is to keep *anyone* who could affect the outcome or the measurement of the response from knowing which subjects have been assigned to which treatments. So, not only should your cola-tasting subjects be blinded, but also *you*, as the experimenter, shouldn't know which drink is which, either—at least until you're ready to analyze the results.

There are two main classes of individuals who can affect the outcome of the experiment:

- ▶ those who could influence the results (the subjects, treatment administrators, or technicians)
- ▶ those who evaluate the results (judges, treating physicians, etc.)

When all the individuals in either one of these classes are blinded, an experiment is said to be **single-blind**. When everyone in *both* classes is blinded, we call the experiment **double-blind**. Even if several individuals in one class are blinded—for example, both the patients and the technicians who administer the treatment—the study would still be just single-blind. If only some of the individuals in a class are blind—for example, if subjects are not told of their treatment, but the administering technician is not

blind—there is a substantial risk that subjects can discern their treatment from subtle cues in the technician's behavior or that the technician might inadvertently treat subjects differently. Such experiments cannot be considered truly blind.

In our tomato experiment, we certainly don't want the people judging the taste to know which tomatoes got the fertilizer. That makes the experiment single-blind. We might also not want the people caring for the tomatoes to know which ones were being fertilized, in case they might treat them differently in other ways, too. We can accomplish this double-blinding by having some fake fertilizer for them to put on the other plants. Read on.

## FOR EXAMPLE

### Blinding

**Recap:** In our experiment to see if the new pet food is now safe, we're feeding one group of dogs the new food and another group a food we know to be safe. Our response variable is the health of the animals as assessed by a veterinarian.

**Questions:** Should the vet be blinded? Why or why not? How would you do this? (Extra credit: Can this experiment be double-blind? Would that mean that the test animals wouldn't know what they were eating?)

*Whenever the response variable involves judgment, it is a good idea to blind the evaluator to the treatments. The veterinarian should not be told which dogs ate which foods.*

*Extra credit: There is a need for double-blinding. In this case, the workers who care for and feed the animals should not be aware of which dogs are receiving which food. We'll need to make the "safe" food look as much like the "test" food as possible.*

## Placebos

**AS** **Activity: Blinded Experiments.** This narrated account of blinding isn't a placebo!

Often, simply applying *any* treatment can induce an improvement. Every parent knows the medicinal value of a kiss to make a toddler's scrape or bump stop hurting. Some of the improvement seen with a treatment—even an effective treatment—can be due simply to the act of treating. To separate these two effects, we can use a control treatment that mimics the treatment itself.

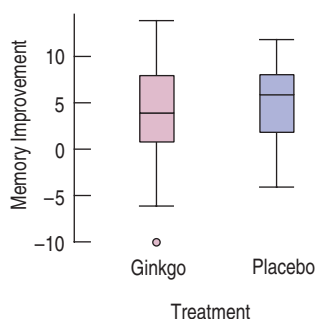
The placebo effect is stronger when placebo treatments are administered with authority or by a figure who appears to be an authority. “Doctors” in white coats generate a stronger effect than salespeople in polyester suits. But the placebo effect is not reduced much even when subjects know that the effect exists. People often suspect that they’ve gotten the placebo if nothing at all happens. So, recently, drug manufacturers have gone so far in making placebos realistic that they cause the same side effects as the drug being tested! Such “active placebos” usually induce a stronger placebo effect. When those side effects include loss of appetite or hair, the practice may raise ethical questions.

A “fake” treatment that looks just like the treatments being tested is called a **placebo**. Placebos are the best way to blind subjects from knowing whether they are receiving the treatment or not. One common version of a placebo in drug testing is a “sugar pill.” Especially when psychological attitude can affect the results, control group subjects treated with a placebo may show an improvement.

The fact is that subjects treated with a placebo sometimes improve. It’s not unusual for 20% or more of subjects given a placebo to report reduction in pain, improved movement, or greater alertness, or even to demonstrate improved health or performance. This **placebo effect** highlights both the importance of effective blinding and the importance of comparing treatments with a control. Placebo controls are so effective that you should use them as an essential tool for blinding whenever possible.

The best experiments are usually

- ▶ randomized.
- ▶ double-blind.
- ▶ comparative.
- ▶ placebo-controlled.



**Does ginkgo biloba improve memory?** Researchers investigated the purported memory-enhancing effect of ginkgo biloba tree extract (P. R. Solomon, F. Adams, A. Silver, J. Zimmer, R. De Veaux, “Ginkgo for Memory Enhancement. A Randomized Controlled Trial.” *JAMA* 288 [2002]: 835–840). In a randomized, comparative, double-blind, placebo-controlled study, they administered treatments to 230 elderly community members. One group received Ginkoba™ according to the manufacturer’s instructions. The other received a similar-looking placebo. Thirteen different tests of memory were administered before and after treatment. The placebo group showed greater improvement on 7 of the tests, the treatment group on the other 6. None showed any significant differences. Here are boxplots of one measure.



By permission of John L. Hart FLP and Creators Syndicate, Inc.

## Blocking

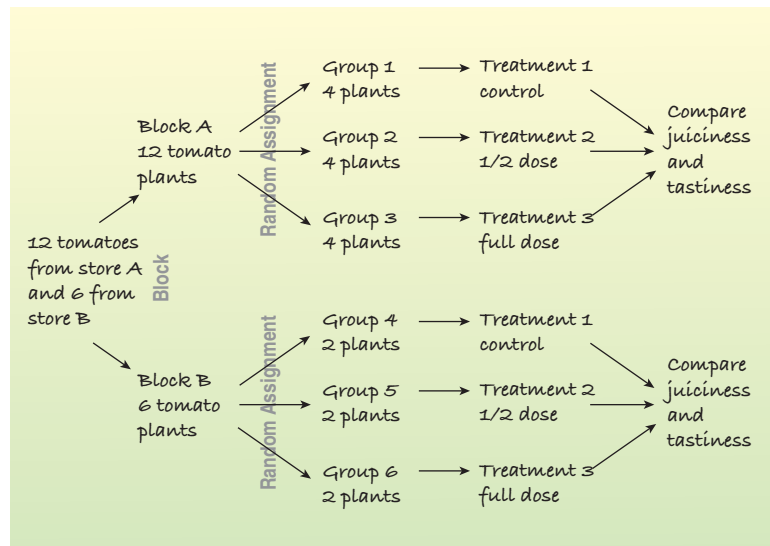
We wanted to use 18 tomato plants of the same variety for our experiment, but suppose the garden store had only 12 plants left. So we drove down to the nursery and bought 6 more plants of that variety. We worry that the tomato plants from the two stores are different somehow, and, in fact, they don’t really look the same.

How can we design the experiment so that the differences between the stores don’t mess up our attempts to see differences among fertilizer levels? We can’t measure the effect of a store the same way as we can the fertilizer because we can’t assign it as we would a factor in the experiment. You can’t tell a tomato what store to come from.

Because stores may vary in the care they give plants or in the sources of their seeds, the plants from either store are likely to be more like each other than they are like the plants from the other store. When groups of experimental units are similar, it's often a good idea to gather them together into **blocks**. By blocking, we isolate the variability attributable to the differences between the blocks, so that we can see the differences caused by the treatments more clearly. Here, we would define the plants from each store to be a block. The randomization is introduced when we randomly assign treatments within each block.

In a completely randomized design, each of the 18 plants would have an equal chance to land in each of the three treatment groups. But we realize that the store may have an effect. To isolate the store effect, we block on store by assigning the plants from each store to treatments at random. So we now have six treatment groups, three for each block. Within each block, we'll randomly assign the same number of plants to each of the three treatments. The experiment is still fair because each treatment is still applied (at random) to the same number of plants and to the same proportion from each store: 4 from store A and 2 from store B. Because the randomization occurs only within the blocks (plants from one store cannot be assigned to treatment groups for the other), we call this a **randomized block design**.

In effect, we conduct two parallel experiments, one for tomatoes from each store, and then combine the results. The picture tells the story:



In a retrospective or prospective study, subjects are sometimes paired because they are similar in ways *not* under study. **Matching** subjects in this way can reduce variation in much the same way as blocking. For example, a retrospective study of music education and grades might match each student who studies an instrument with someone of the same sex who is similar in family income but didn't study an instrument. When we compare grades of music students with those of non-music students, the matching would reduce the variation due to income and sex differences.

Blocking is the same idea for experiments as stratifying is for sampling. Both methods group together subjects that are similar and randomize within those groups as a way to remove unwanted variation. (But be careful to keep the terms straight. Don't say that we "stratify" an experiment or "block" a sample.) We use blocks to reduce variability so we can see the effects of the factors; we're not usually interested in studying the effects of the blocks themselves.

## FOR EXAMPLE

## Blocking

**Recap:** In 2007, pet food contamination put cats at risk, as well as dogs. Our experiment should probably test the safety of the new food on both animals.

**Questions:** Why shouldn't we randomly assign a mix of cats and dogs to the two treatment groups? What would you recommend instead?

*Dogs and cats might respond differently to the foods, and that variability could obscure my results. Blocking by species can remove that superfluous variation. I'd randomize cats to the two treatments (test food and safe food) separately from the dogs. I'd measure their responses separately and look at the results afterward.*



## JUST CHECKING

2. Recall the experiment about gastric freezing, an old method for treating peptic ulcers that you read about in the first Just Checking. Doctors would insert an inflatable bladder down the patient's esophagus and into the stomach and then pump in a cold liquid. A major insurance company now states that it doesn't cover this treatment because "double-blind, controlled clinical trials" failed to demonstrate that gastric freezing was effective.
  - a) What does it mean that the experiment was double-blind?
  - b) Why would you recommend a placebo control?
  - c) Suppose that researchers suspected that the effectiveness of the gastric freezing treatment might depend on whether a patient had recently developed the peptic ulcer or had been suffering from the condition for a long time. How might the researchers have designed the experiment?

## Adding More Factors

There are two kinds of gardeners. Some water frequently, making sure that the plants are never dry. Others let Mother Nature take her course and leave the watering to her. The makers of OptiGro want to ensure that their product will work under a wide variety of watering conditions. Maybe we should include the amount of watering as part of our experiment. Can we study a second factor at the same time and still learn as much about fertilizer?

We now have two factors (fertilizer at three levels and irrigation at two levels). We combine them in all possible ways to yield six treatments:

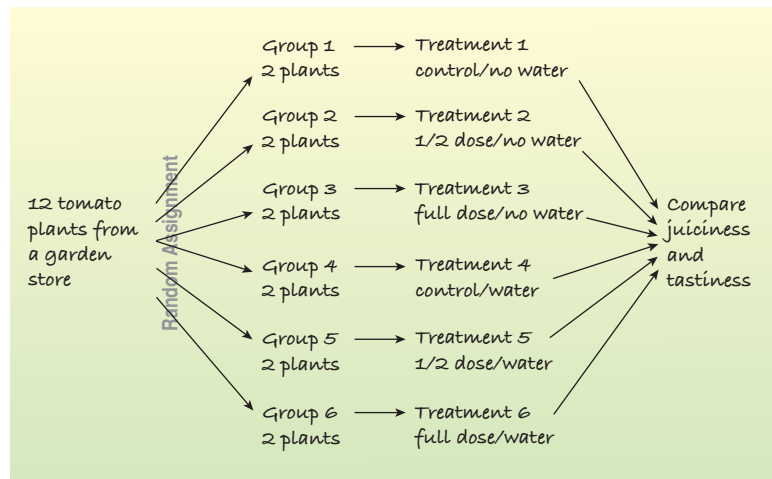
	No Fertilizer	Half Fertilizer	Full Fertilizer
No Added Water	1	2	3
Daily Watering	4	5	6

If we allocate the original 12 plants, the experiment now assigns 2 plants to each of these six treatments at random. This experiment is a **completely randomized two-factor experiment** because any plant could end up assigned at random to any of the six treatments (and we have two factors).

It's often important to include several factors in the same experiment in order to see what happens when the factor levels are applied in different *combinations*. A common misconception is that applying several factors at once makes it difficult to separate the effects of the individual factors. You may hear people say that experiments should always be run "one factor at a time." In fact, just the opposite

### Think Like a Statistician

With two factors, we can account for more of the variation. That lets us see the underlying patterns more clearly.



is true: Experiments with more than one factor are both more efficient and provide more information than one-at-a-time experiments. There are many ways to design efficient multifactor experiments. You can take a whole course on the design and analysis of such experiments.

## Confounding

Professor Stephen Ceci of Cornell University performed an experiment to investigate the effect of a teacher's classroom style on student evaluations. He taught a class in developmental psychology during two successive terms to a total of 472 students in two very similar classes. He kept everything about his teaching identical (same text, same syllabus, same office hours, etc.) and modified only his style in class. During the fall term, he maintained a subdued demeanor. During the spring term, he used expansive gestures and lectured with more enthusiasm, varying his vocal pitch and using more hand gestures. He administered a standard student evaluation form at the end of each term.

The students in the fall term class rated him only an average teacher. Those in the spring term class rated him an excellent teacher, praising his knowledge and accessibility, and even the quality of the textbook. On the question "How much did you learn in the course?" the average response changed from 2.93 to 4.05 on a 5-point scale.<sup>5</sup>

How much of the difference he observed was due to his difference in manner, and how much might have been due to the season of the year? Fall term in Ithaca, NY (home of Cornell University), starts out colorful and pleasantly warm but ends cold and bleak. Spring term starts out bitter and snowy and ends with blooming flowers and singing birds. Might students' overall happiness have been affected by the season and reflected in their evaluations?

Unfortunately, there's no way to tell. Nothing in the data enables us to tease apart these two effects, because all the students who experienced the subdued manner did so during the fall term and all who experienced the expansive manner did so during the spring. When the levels of one factor are associated with the levels of another factor, we say that these two factors are **confounded**.

In some experiments, such as this one, it's just not possible to avoid some confounding. Professor Ceci could have randomly assigned students to one of two classes during the same term, but then we might question whether mornings or

<sup>5</sup> But the two classes performed almost identically well on the final exam.



afternoons were better, or whether he really delivered the same class the second time (after practicing on the first class). Or he could have had another professor deliver the second class, but that would have raised more serious issues about differences in the two professors and concern over more serious confounding.

## FOR EXAMPLE

### Confounding

**Recap:** After many dogs and cats suffered health problems caused by contaminated foods, we're trying to find out whether a newly formulated pet food is safe. Our experiment will feed some animals the new food and others a food known to be safe, and a veterinarian will check the response.

**Question:** Why would it be a bad design to feed the test food to some dogs and the safe food to cats?

*This would create confounding. We would not be able to tell whether any differences in animals' health were attributable to the food they had eaten or to differences in how the two species responded.*



**A two-factor example** Confounding can also arise from a badly designed multifactor experiment. Here's a classic. A credit card bank wanted to test the sensitivity of the market to two factors: the annual fee charged for a card and the annual percentage rate charged. Not wanting to scrimp on sample size, the bank selected 100,000 people at random from a mailing list. It sent out 50,000 offers with a low rate and no fee and 50,000 offers with a higher rate and a \$50 annual fee. Guess what happened? That's right—people preferred the low-rate, no-fee card. No surprise. In fact, they signed up for that card at over twice the rate as the other offer. And because of the large sample size, the bank was able to estimate the difference precisely. But the question the bank really wanted to answer was “how much of the change was due to the rate, and how much was due to the fee?” unfortunately, there's simply no way to separate out the two effects. If the bank had sent out all four possible different treatments—low rate with no fee, low rate with \$50 fee, high rate with no fee, and high rate with \$50 fee—each to 25,000 people, it could have learned about both factors and could have also seen what happens when the two factors occur in combination.

## Lurking or Confounding?

Confounding may remind you of the problem of lurking variables we discussed back in Chapters 7 and 9. Confounding variables and lurking variables are alike in that they interfere with our ability to interpret our analyses simply. Each can mislead us, but there are important differences in both how and where the confusion may arise.

A lurking variable creates an association between two other variables that tempts us to think that one may cause the other. This can happen in a regression analysis or an observational study when a lurking variable influences both the explanatory and response variables. Recall that countries with more TV sets per capita tend to have longer life expectancies. We shouldn't conclude it's the TVs “causing” longer life. We suspect instead that a generally higher standard of living may mean that people can afford more TVs and get better health care, too. Our data revealed an association between TVs and life expectancy, but economic conditions were a likely lurking variable. A lurking variable, then, is usually thought of as a variable associated with both  $y$  and  $x$  that makes it appear that  $x$  may be causing  $y$ .

Confounding can arise in experiments when some other variable associated with a factor has an effect on the response variable. However, in a designed experiment, the experimenter *assigns* treatments (at random) to subjects rather than just observing them. A confounding variable can't be thought of as causing that assignment. Professor Ceci's choice of teaching styles was not caused by the weather, but because he used one style in the fall and the other in spring, he was unable to tell how much of his students' reactions were attributable to his teaching and how much to the weather. A confounding variable, then, is associated in a noncausal way with a factor and affects the response. Because of the confounding, we find that we can't tell whether any effect we see was caused by our factor or by the confounding variable—or even by both working together.

Both confounding and lurking variables are outside influences that make it harder to understand the relationship we are modeling. However, the nature of the causation is different in the two situations. In regression and observational studies, we can only observe associations between variables. Although we can't demonstrate a causal relationship, we often imagine whether  $x$  could cause  $y$ . We can be misled by a lurking variable that influences both. In a designed experiment, we often hope to show that the factor causes a response. Here we can be misled by a confounding variable that's associated with the factor and causes or contributes to the differences we observe in the response.

It's worth noting that the role of blinding in an experiment is to combat a possible source of confounding. There's a risk that knowledge about the treatments could lead the subjects or those interacting with them to behave differently or could influence judgments made by the people evaluating the responses. That means we won't know whether the treatments really do produce different results or if we're being fooled by these confounding influences.


## WHAT CAN GO WRONG?

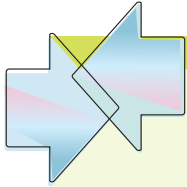
- ▶ **Don't give up just because you can't run an experiment.** Sometimes we can't run an experiment because we can't identify or control the factors. Sometimes it would simply be unethical to run the experiment. (Consider randomly assigning students to take—and be graded in—a Statistics course deliberately taught to be boring and difficult or one that had an unlimited budget to use multimedia, real-world examples, and field trips to make the subject more interesting.) If we can't perform an experiment, often an observational study is a good choice.
- ▶ **Beware of confounding.** Use randomization whenever possible to ensure that the factors not in your experiment are not confounded with your treatment levels. Be alert to confounding that cannot be avoided, and report it along with your results.
- ▶ **Bad things can happen even to good experiments.** Protect yourself by recording additional information. An experiment in which the air conditioning failed for 2 weeks, affecting the results, was saved by recording the temperature (although that was not originally one of the factors) and estimating the effect the higher temperature had on the response.<sup>6</sup>

It's generally good practice to collect as much information as possible about your experimental units and the circumstances of the experiment. For example, in the tomato experiment, it would be wise to record details of the weather (temperature, rainfall, sunlight) that might affect the plants and any facts available about their

<sup>6</sup> R. D. DeVeaux and M. Szelewski, "Optimizing Automatic Splitless Injection Parameters for Gas Chromatographic Environmental Analysis." *Journal of Chromatographic Science* 27, no. 9 (1989): 513–518.

growing situation. (Is one side of the field in shade sooner than the other as the day proceeds? Is one area lower and a bit wetter?) Sometimes we can use this extra information during the analysis to reduce biases.

- ▶ **Don't spend your entire budget on the first run.** Just as it's a good idea to pretest a survey, it's always wise to try a small pilot experiment before running the full-scale experiment. You may learn, for example, how to choose factor levels more effectively, about effects you forgot to control, and about unanticipated confoundings. 



## CONNECTIONS

The fundamental role of randomization in experiments clearly points back to our discussions of randomization, to our experiments with simulations, and to our use of randomization in sampling. The similarities and differences between experiments and samples are important to keep in mind and can make each concept clearer.

If you think that blocking in an experiment resembles stratifying in a sample, you're quite right. Both are ways of removing variation we can identify to help us see past the variation in the data.

Experiments compare groups of subjects that have been treated differently. Graphics such as boxplots that help us compare groups are closely related to these ideas. Think about what we look for in a boxplot to tell whether two groups look really different, and you'll be thinking about the same issues as experiment designers.

Generally, we're going to consider how different the mean responses are for different treatment groups. And we're going to judge whether those differences are large by using standard deviations as rulers. (That's why we needed to replicate results for each treatment; we need to be able to estimate those standard deviations.) The discussion in Chapter 6 introduced this fundamental statistical thought, and it's going to keep coming back over and over again. Statistics is about variation.

We'll see a number of ways to analyze results from experiments in subsequent chapters.



## WHAT HAVE WE LEARNED?

We've learned to recognize sample surveys, observational studies, and randomized comparative experiments. We know that these methods collect data in different ways and lead us to different conclusions.

We've learned to identify retrospective and prospective observational studies and understand the advantages and disadvantages of each.

We've learned that only well-designed experiments can allow us to reach cause-and-effect conclusions. We manipulate levels of treatments to see if the factor we have identified produces changes in our response variable.

We've learned the principles of experimental design:

- ▶ We want to be sure that variation in the response variable can be attributed to our factor, so we identify and control as many other sources of variability as possible.
- ▶ Because there are many possible sources of variability that we cannot identify, we try to equalize those by randomly assigning experimental units to treatments.
- ▶ We replicate the experiment on as many subjects as possible.
- ▶ We consider blocking to reduce variability from sources we recognize but cannot control.

We've learned the value of having a control group and of using blinding and placebo controls.

Finally, we've learned to recognize the problems posed by confounding variables in experiments and lurking variables in observational studies.

## Terms

Observational study	292. A study based on data in which no manipulation of factors has been employed.
Retrospective study	292. An observational study in which subjects are selected and then their previous conditions or behaviors are determined. Retrospective studies need not be based on random samples and they usually focus on estimating differences between groups or associations between variables.
Prospective study	293. An observational study in which subjects are followed to observe future outcomes. Because no treatments are deliberately applied, a prospective study is not an experiment. Nevertheless, prospective studies typically focus on estimating differences among groups that might appear as the groups are followed during the course of the study.
Experiment	294. An experiment <i>manipulates</i> factor levels to create treatments, <i>randomly assigns</i> subjects to these treatment levels, and then <i>compares</i> the responses of the subject groups across treatment levels.
Random assignment	294. To be valid, an experiment must assign experimental units to treatment groups at random. This is called random assignment.
Factor	294. A variable whose levels are manipulated by the experimenter. Experiments attempt to discover the effects that differences in factor levels may have on the responses of the experimental units.
Response	294. A variable whose values are compared across different treatments. In a randomized experiment, large response differences can be attributed to the effect of differences in treatment level.
Experimental units	294. Individuals on whom an experiment is performed. Usually called <b>subjects</b> or <b>participants</b> when they are human.
Level	294. The specific values that the experimenter chooses for a factor are called the levels of the factor.
Treatment	294. The process, intervention, or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.
Principles of experimental design	<ul style="list-style-type: none"> <li>▶ 295. <b>Control</b> aspects of the experiment that we know may have an effect on the response, but that are not the factors being studied.</li> <li>▶ 296. <b>Randomize</b> subjects to treatments to even out effects that we cannot control.</li> <li>▶ 296. <b>Replicate</b> over as many subjects as possible. Results for a single subject are just anecdotes. If, as often happens, the subjects of the experiment are not a representative sample from the population of interest, replicate the entire study with a different group of subjects, preferably from a different part of the population.</li> <li>▶ 296. <b>Block</b> to reduce the effects of identifiable attributes of the subjects that cannot be controlled.</li> </ul>
Statistically significant	299. When an observed difference is too large for us to believe that it is likely to have occurred naturally, we consider the difference to be statistically significant. Subsequent chapters will show specific calculations and give rules, but the principle remains the same.
Control group	301. The experimental units assigned to a baseline treatment level, typically either the default treatment, which is well understood, or a null, placebo treatment. Their responses provide a basis for comparison.
Blinding	301. Any individual associated with an experiment who is not aware of how subjects have been allocated to treatment groups is said to be blinded.
Single-blind	302. There are two main classes of individuals who can affect the outcome of an experiment:
Double-blind	<ul style="list-style-type: none"> <li>▶ those who could <i>influence the results</i> (the subjects, treatment administrators, or technicians).</li> <li>▶ those who <i>evaluate the results</i> (judges, treating physicians, etc.).</li> </ul> <p>When every individual in <i>either</i> of these classes is blinded, an experiment is said to be single-blind. When everyone in <i>both</i> classes is blinded, we call the experiment double-blind.</p>
Placebo	303. A treatment known to have no effect, administered so that all groups experience the same conditions. Many subjects respond to such a treatment (a response known as a placebo effect). Only by comparing with a placebo can we be sure that the observed effect of a treatment is not due simply to the placebo effect.
Placebo effect	303. The tendency of many human subjects (often 20% or more of experiment subjects) to show a response even when administered a placebo.

**Blocking** 303. When groups of experimental units are similar, it is often a good idea to gather them together into blocks. By blocking, we isolate the variability attributable to the differences between the blocks so that we can see the differences caused by the treatments more clearly.

**Matching** 304. In a retrospective or prospective study, subjects who are similar in ways not under study may be matched and then compared with each other on the variables of interest. Matching, like blocking, reduces unwanted variation.

**Designs** 298, 305. In a **completely randomized design**, all experimental units have an equal chance of receiving any treatment.

304. In a **randomized block design**, the randomization occurs only within blocks.

**Confounding** 306. When the levels of one factor are associated with the levels of another factor in such a way that their effects cannot be separated, we say that these two factors are confounded.

## Skills

### THINK

- ▶ Recognize when an observational study would be appropriate.
- ▶ Be able to identify observational studies as retrospective or prospective, and understand the strengths and weaknesses of each method.
- ▶ Know the four basic principles of sound experimental design—control, randomize, replicate, and block—and be able to explain each.
- ▶ Be able to recognize the factors, the treatments, and the response variable in a description of a designed experiment.
- ▶ Understand the essential importance of randomization in assigning treatments to experimental units.
- ▶ Understand the importance of replication to move from anecdotes to general conclusions.
- ▶ Understand the value of blocking so that variability due to differences in attributes of the subjects can be removed.
- ▶ Understand the importance of a control group and the need for a placebo treatment in some studies.
- ▶ Understand the importance of blinding and double-blinding in studies on human subjects, and be able to identify blinding and the need for blinding in experiments.
- ▶ Understand the value of a placebo in experiments with human participants.

### SHOW

- ▶ Be able to design a completely randomized experiment to test the effect of a single factor.
- ▶ Be able to design an experiment in which blocking is used to reduce variation.
- ▶ Know how to use graphical displays to compare responses for different treatment groups. Understand that you should *never* proceed with any other analysis of a designed experiment without first looking at boxplots or other graphical displays.

### TELL

- ▶ Know how to report the results of an observational study. Identify the subjects, how the data were gathered, and any potential biases or flaws you may be aware of. Identify the factors known and those that might have been revealed by the study.
- ▶ Know how to compare the responses in different treatment groups to assess whether the differences are larger than could be reasonably expected from ordinary sampling variability.
- ▶ Know how to report the results of an experiment. Tell who the subjects are and how their assignment to treatments was determined. Report how and in what measurement units the response variable was measured.
- ▶ Understand that your description of an experiment should be sufficient for another researcher to replicate the study with the same methods.
- ▶ Be able to report on the statistical significance of the result in terms of whether the observed group-to-group differences are larger than could be expected from ordinary sampling variation.

## EXPERIMENTS ON THE COMPUTER

Most experiments are analyzed with a statistics package. You should almost always display the results of a comparative experiment with side-by-side boxplots. You may also want to display the means and standard deviations of the treatment groups in a table.

The analyses offered by statistics packages for comparative randomized experiments fall under the general heading of Analysis of Variance, usually abbreviated ANOVA. These analyses are beyond the scope of this chapter.

## EXERCISES

- Standardized test scores.** For his Statistics class experiment, researcher J. Gilbert decided to study how parents' income affects children's performance on standardized tests like the SAT. He proposed to collect information from a random sample of test takers and examine the relationship between parental income and SAT score.
    - Is this an experiment? If not, what kind of study is it?
    - If there is relationship between parental income and SAT score, why can't we conclude that differences in score are caused by differences in parental income?
  - Heart attacks and height.** Researchers who examined health records of thousands of males found that men who died of myocardial infarction (heart attack) tended to be shorter than men who did not.
    - Is this an experiment? If not, what kind of study is it?
    - Is it correct to conclude that shorter men are at higher risk for heart attack? Explain.
  - MS and vitamin D.** Multiple sclerosis (MS) is an autoimmune disease that strikes more often the farther people live from the equator. Could vitamin D—which most people get from the sun's ultraviolet rays—be a factor? Researchers compared vitamin D levels in blood samples from 150 U.S. military personnel who have developed MS with blood samples of nearly 300 who have not. The samples were taken, on average, five years before the disease was diagnosed. Those with the highest blood vitamin D levels had a 62% lower risk of MS than those with the lowest levels. (The link was only in whites, not in blacks or Hispanics.)
    - What kind of study was this?
    - Is that an appropriate choice for investigating this problem? Explain.
    - Who were the subjects?
    - What were the variables?
  - Super Bowl commercials.** When spending large amounts to purchase advertising time, companies want to know what audience they'll reach. In January 2007, a poll asked 1008 American adults whether they planned to watch the upcoming Super Bowl. Men and women were asked separately whether they were looking forward more to the football game or to watching the commercials. Among the men, 16% were planning to watch and were looking forward primarily to the commercials. Among women, 30% were looking forward primarily to the commercials.
    - Was this a stratified sample or a blocked experiment? Explain.
    - Was the design of the study appropriate for the advertisers' questions?
  - Menopause.** Researchers studied the herb black cohosh as a treatment for hot flashes caused by menopause. They randomly assigned 351 women aged 45 to 55 who reported at least two hot flashes a day to one of five groups: (1) black cohosh, (2) a multiherb supplement with black cohosh, (3) the multiherb supplement plus advice to consume more soy foods, (4) estrogen replacement therapy, or (5) receive a placebo. After a year, only the women given estrogen replacement therapy had symptoms different from those of the placebo group. [*Annals of Internal Medicine* 145:12, 869–897]
    - What kind of study was this?
    - Is that an appropriate choice for this problem?
    - Who were the subjects?
    - Identify the treatment and response variables.
  - Honesty.** Coffee stations in offices often just ask users to leave money in a tray to pay for their coffee, but many people cheat. Researchers at Newcastle University replaced the picture of flowers on the wall behind the coffee station with a picture of staring eyes. They found that the average contribution increased significantly above the well-established standard when people felt they were being watched, even though the eyes were patently not real. (*NY Times* 12/10/06)
    - Was this a survey, an observational study, or an experiment? How can we tell?
    - Identify the variables.
    - What does "increased significantly" mean in a statistical sense?
- 7–20. **What's the design?** Read each brief report of statistical research, and identify
- whether it was an observational study or an experiment. If it was an observational study, identify (if possible)
  - whether it was retrospective or prospective.
  - the subjects studied and how they were selected.

- d) the parameter of interest.
- e) the nature and scope of the conclusion the study can reach.

*If it was an experiment, identify (if possible)*

- b) the subjects studied.
  - c) the factor(s) in the experiment and the number of levels for each.
  - d) the number of treatments.
  - e) the response variable measured.
  - f) the design (completely randomized, blocked, or matched).
  - g) whether it was blind (or double-blind).
  - h) the nature and scope of the conclusion the experiment can reach.
7. Over a 4-month period, among 30 people with bipolar disorder, patients who were given a high dose (10 g/day) of omega-3 fats from fish oil improved more than those given a placebo. (*Archives of General Psychiatry* 56 [1999]: 407)
  8. Among a group of disabled women aged 65 and older who were tracked for several years, those who had a vitamin B<sub>12</sub> deficiency were twice as likely to suffer severe depression as those who did not. (*American Journal of Psychiatry* 157 [2000]: 715)
  9. In a test of roughly 200 men and women, those with moderately high blood pressure (averaging 164/89 mm Hg) did worse on tests of memory and reaction time than those with normal blood pressure. (*Hypertension* 36 [2000]: 1079)
  10. Is diet or exercise effective in combating insomnia? Some believe that cutting out desserts can help alleviate the problem, while others recommend exercise. Forty volunteers suffering from insomnia agreed to participate in a month-long test. Half were randomly assigned to a special no-desserts diet; the others continued desserts as usual. Half of the people in each of these groups were randomly assigned to an exercise program, while the others did not exercise. Those who ate no desserts and engaged in exercise showed the most improvement.
  11. After menopause, some women take supplemental estrogen. There is some concern that if these women also drink alcohol, their estrogen levels will rise too high. Twelve volunteers who were receiving supplemental estrogen were randomly divided into two groups, as were 12 other volunteers not on estrogen. In each case, one group drank an alcoholic beverage, the other a nonalcoholic beverage. An hour later, everyone's estrogen level was checked. Only those on supplemental estrogen who drank alcohol showed a marked increase.
  12. Researchers have linked an increase in the incidence of breast cancer in Italy to dioxin released by an industrial accident in 1976. The study identified 981 women who lived near the site of the accident and were under age 40 at the time. Fifteen of the women had developed breast cancer at an unusually young average age of 45. Medical records showed that they had heightened concentrations of dioxin in their blood and that each tenfold increase in dioxin level was associated with a doubling of the risk of breast cancer. (*Science News*, Aug. 3, 2002)
  13. In 2002 the journal *Science* reported that a study of women in Finland indicated that having sons shortened the life-spans of mothers by about 34 weeks per son, but that daughters helped to lengthen the mothers' lives. The data came from church records from the period 1640 to 1870.
  14. Scientists at a major pharmaceutical firm investigated the effectiveness of an herbal compound to treat the common cold. They exposed each subject to a cold virus, then gave him or her either the herbal compound or a sugar solution known to have no effect on colds. Several days later they assessed the patient's condition, using a cold severity scale ranging from 0 to 5. They found no evidence of benefits associated with the compound.
  15. The May 4, 2000, issue of *Science News* reported that, contrary to popular belief, depressed individuals cry no more often in response to sad situations than nondepressed people. Researchers studied 23 men and 48 women with major depression and 9 men and 24 women with no depression. They showed the subjects a sad film about a boy whose father has died, noting whether or not the subjects cried. Women cried more often than men, but there were no significant differences between the depressed and nondepressed groups.
  16. Some people who race greyhounds give the dogs large doses of vitamin C in the belief that the dogs will run faster. Investigators at the University of Florida tried three different diets in random order on each of five racing greyhounds. They were surprised to find that when the dogs ate high amounts of vitamin C they ran more slowly. (*Science News*, July 20, 2002)
  17. Some people claim they can get relief from migraine headache pain by drinking a large glass of ice water. Researchers plan to enlist several people who suffer from migraines in a test. When a participant experiences a migraine headache, he or she will take a pill that may be a standard pain reliever or a placebo. Half of each group will also drink ice water. Participants will then report the level of pain relief they experience.
  18. A dog food company wants to compare a new lower-calorie food with their standard dog food to see if it's effective in helping inactive dogs maintain a healthy weight. They have found several dog owners willing to participate in the trial. The dogs have been classified as small, medium, or large breeds, and the company will supply some owners of each size of dog with one of the two foods. The owners have agreed not to feed their dogs anything else for a period of 6 months, after which the dogs' weights will be checked.
  19. Athletes who had suffered hamstring injuries were randomly assigned to one of two exercise programs. Those who engaged in static stretching returned to sports activity in a mean of 15.2 days faster than those assigned to a program of agility and trunk stabilization exercises. (*Journal of Orthopaedic & Sports Physical Therapy* 34 [March 2004]: 3)
  20. Pew Research compared respondents to an ordinary 5-day telephone survey with respondents to a 4-month-long rigorous survey designed to generate the highest

possible response rate. They were especially interested in identifying any variables for which those who responded to the ordinary survey were different from those who could be reached only by the rigorous survey.

21. **Omega-3.** Exercise 7 describes an experiment that showed that high doses of omega-3 fats might be of benefit to people with bipolar disorder. The experiment involved a control group of subjects who received a placebo. Why didn't the experimenters just give everyone the omega-3 fats to see if they improved?
22. **Insomnia.** Exercise 10 describes an experiment showing that exercise helped people sleep better. The experiment involved other groups of subjects who didn't exercise. Why didn't the experimenters just have everyone exercise and see if their ability to sleep improved?
23. **Omega-3 revisited.** Exercises 7 and 21 describe an experiment investigating a dietary approach to treating bipolar disorder. Researchers randomly assigned 30 subjects to two treatment groups, one group taking a high dose of omega-3 fats and the other a placebo.
  - a) Why was it important to randomize in assigning the subjects to the two groups?
  - b) What would be the advantages and disadvantages of using 100 subjects instead of 30?
24. **Insomnia again.** Exercises 10 and 22 describe an experiment investigating the effectiveness of exercise in combating insomnia. Researchers randomly assigned half of the 40 volunteers to an exercise program.
  - a) Why was it important to randomize in deciding who would exercise?
  - b) What would be the advantages and disadvantages of using 100 subjects instead of 40?
25. **Omega-3, finis.** Exercises 7, 21, and 23 describe an experiment investigating the effectiveness of omega-3 fats in treating bipolar disorder. Suppose some of the 30 subjects were very active people who walked a lot or got vigorous exercise several times a week, while others tended to be more sedentary, working office jobs and watching a lot of TV. Why might researchers choose to block the subjects by activity level before randomly assigning them to the omega-3 and placebo groups?
26. **Insomnia, at last.** Exercises 10, 22, and 24 describe an experiment investigating the effectiveness of exercise in combating insomnia. Suppose some of the 40 subjects had maintained a healthy weight, but others were quite overweight. Why might researchers choose to block the subjects by weight level before randomly assigning some of each group to the exercise program?
27. **Tomatoes.** Describe a strategy to randomly split the 24 tomato plants into the three groups for the chapter's completely randomized single factor test of OptiGro fertilizer.
28. **Tomatoes II.** The chapter also described a completely randomized two-factor experiment testing OptiGro fertilizer in conjunction with two different routines for watering the plants. Describe a strategy to randomly assign the 24 tomato plants to the six treatments.
29. **Shoes.** A running-shoe manufacturer wants to test the effect of its new sprinting shoe on 100-meter dash times.
 

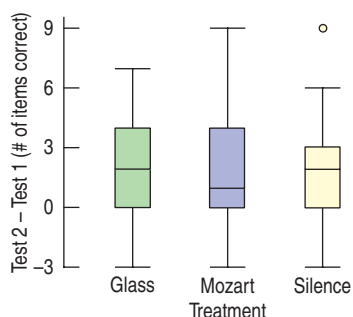
The company sponsors 5 athletes who are running the 100-meter dash in the 2004 Summer Olympic games. To test the shoe, it has all 5 runners run the 100-meter dash with a competitor's shoe and then again with their new shoe. The company uses the difference in times as the response variable.

  - a) Suggest some improvements to the design.
  - b) Why might the shoe manufacturer not be able to generalize the results they find to all runners?
30. **Swimsuits.** A swimsuit manufacturer wants to test the speed of its newly designed suit. The company designs an experiment by having 6 randomly selected Olympic swimmers swim as fast as they can with their old swimsuit first and then swim the same event again with the new, expensive swimsuit. The company will use the difference in times as the response variable. Criticize the experiment and point out some of the problems with generalizing the results.
31. **Hamstrings.** Exercise 19 discussed an experiment to see if the time it took athletes with hamstring injuries to be able to return to sports was different depending on which of two exercise programs they engaged in.
  - a) Explain why it was important to assign the athletes to the two different treatments randomly.
  - b) There was no control group consisting of athletes who did not participate in a special exercise program. Explain the advantage of including such a group.
  - c) How might blinding have been used?
  - d) One group returned to sports activity in a mean of 37.4 days ( $SD = 27.6$  days) and the other in a mean of 22.2 days ( $SD = 8.3$  days). Do you think this difference is statistically significant? Explain.
32. **Diet and blood pressure.** An experiment that showed that subjects fed the DASH diet were able to lower their blood pressure by an average of 6.7 points compared to a group fed a "control diet." All meals were prepared by dietitians.
  - a) Why were the subjects randomly assigned to the diets instead of letting people pick what they wanted to eat?
  - b) Why were the meals prepared by dietitians?
  - c) Why did the researchers need the control group? If the DASH diet group's blood pressure was lower at the end of the experiment than at the beginning, wouldn't that prove the effectiveness of that diet?
  - d) What additional information would you want to know in order to decide whether an average reduction in blood pressure of 6.7 points was statistically significant?
33. **Mozart.** Will listening to a Mozart piano sonata make you smarter? In a 1995 study published in the journal *Psychological Science*, Rauscher, Shaw, and Ky reported that when students were given a spatial reasoning section of a standard IQ test, those who listened to Mozart for 10 minutes improved their scores more than those who simply sat quietly.
  - a) These researchers said the differences were statistically significant. Explain what that means in context.
  - b) Steele, Bass, and Crook tried to replicate the original study. In their study, also published in *Psychological Science* (1999), the subjects were 125 college students



who participated in the experiment for course credit. Subjects first took the test. Then they were assigned to one of three groups: listening to a Mozart piano sonata, listening to music by Philip Glass, and sitting for 10 minutes in silence. Three days after the treatments, they were retested. Draw a diagram displaying the design of this experiment.

- c) These boxplots show the differences in score before and after treatment for the three groups. Did the Mozart group show improvement?



- d) Do you think the results prove that listening to Mozart is beneficial? Explain.

34. **Full moon.** It's a common belief that people behave strangely when there's a full moon and that as a result police and emergency rooms are busier than usual. Design a way you could find out whether there is any merit to this belief. Will you use an observational study or an experiment? Why?
35. **Wine.** A 2001 Danish study published in the *Archives of Internal Medicine* casts significant doubt on suggestions that adults who drink wine have higher levels of "good" cholesterol and fewer heart attacks. These researchers followed a group of individuals born at a Copenhagen hospital between 1959 and 1961 for 40 years. Their study found that in this group the adults who drank wine were richer and better educated than those who did not.
- What kind of study was this?
  - It is generally true that people with high levels of education and high socioeconomic status are healthier than others. How does this call into question the supposed health benefits of wine?
  - Can studies such as these prove causation (that wine helps prevent heart attacks, that drinking wine makes one richer, that being rich helps prevent heart attacks, etc.)? Explain.
36. **Swimming.** Recently, a group of adults who swim regularly for exercise were evaluated for depression. It turned out that these swimmers were less likely to be depressed than the general population. The researchers said the difference was statistically significant.
- What does "statistically significant" mean in this context?
  - Is this an experiment or an observational study? Explain.
  - News reports claimed this study proved that swimming can prevent depression. Explain why this conclusion is not justified by the study. Include an example of a possible lurking variable.
- d) But perhaps it is true. We wonder if exercise can ward off depression, and whether anaerobic exercise (like weight training) is as effective as aerobic exercise (like swimming). We find 120 volunteers not currently engaged in a regular program of exercise. Design an appropriate experiment.
37. **Dowsing.** Before drilling for water, many rural homeowners hire a dowser (a person who claims to be able to sense the presence of underground water using a forked stick.) Suppose we wish to set up an experiment to test one dowser's ability. We get 20 identical containers, fill some with water, and ask him to tell which ones they are.
- How will we randomize this procedure?
  - The dowser correctly identifies the contents of 12 out of 20 containers. Do you think this level of success is statistically significant? Explain.
  - How many correct identifications (out of 20) would the dowser have to make to convince you that the forked-stick trick works? Explain.
38. **Healing.** A medical researcher suspects that giving post-surgical patients large doses of vitamin E will speed their recovery times by helping their incisions heal more quickly. Design an experiment to test this conjecture. Be sure to identify the factors, levels, treatments, response variable, and the role of randomization.
39. **Reading.** Some schools teach reading using phonics (the sounds made by letters) and others using whole language (word recognition). Suppose a school district wants to know which method works better. Suggest a design for an appropriate experiment.
40. **Gas mileage.** Do cars get better gas mileage with premium instead of regular unleaded gasoline? It might be possible to test some engines in a laboratory, but we'd rather use real cars and real drivers in real day-to-day driving, so we get 20 volunteers. Design the experiment.
41. **Weekend deaths.** A study published in the *New England Journal of Medicine* (Aug. 2001) suggests that it's dangerous to enter a hospital on a weekend. During a 10-year period, researchers tracked over 4 million emergency admissions to hospitals in Ontario, Canada. Their findings revealed that patients admitted on weekends had a much higher risk of death than those who went on weekdays.
- The researchers said the difference in death rates was "statistically significant." Explain in this context what that means.
  - What kind of study was this? Explain.
  - If you think you're quite ill on a Saturday, should you wait until Monday to seek medical help? Explain.
  - Suggest some possible explanations for this troubling finding.
42. **Shingles.** A research doctor has discovered a new ointment that she believes will be more effective than the current medication in the treatment of shingles (a painful skin rash). Eight patients have volunteered to participate in the initial trials of this ointment. You are the statistician hired as a consultant to help design a completely randomized experiment.
- Describe how you will conduct this experiment.
  - Suppose the eight patients' last names start with the letters A to H. Using the random numbers listed below,

show which patients you will assign to each treatment. Explain your randomization procedure clearly.

41098 18329 78458 31685 55259

- c) Can you make this experiment double-blind? How?
- d) The initial experiment revealed that males and females may respond differently to the ointment. Further testing of the drug's effectiveness is now planned, and many patients have volunteered. What changes in your first design, if any, would you make for this second stage of testing?
43. **Beetles.** Hoping to learn how to control crop damage by a certain species of beetle, a researcher plans to test two different pesticides in small plots of corn. A few days after application of the chemicals, he'll check the number of beetle larvae found on each plant. The researcher wants to know whether either pesticide works and whether there is a significant difference in effectiveness between them. Design an appropriate experiment.
44. **SAT Prep.** Can special study courses actually help raise SAT scores? One organization says that the 30 students they tutored achieved an average gain of 60 points when they retook the test.
- a) Explain why this does not necessarily prove that the special course caused the scores to go up.
- b) Propose a design for an experiment that could test the effectiveness of the tutorial course.
- c) Suppose you suspect that the tutorial course might be more helpful for students whose initial scores were particularly low. How would this affect your proposed design?
45. **Safety switch.** An industrial machine requires an emergency shutoff switch that must be designed so that it can be easily operated with either hand. Design an experiment to find out whether workers will be able to deactivate the machine as quickly with their left hands as with their right hands. Be sure to explain the role of randomization in your design.
46. **Washing clothes.** A consumer group wants to test the effectiveness of a new "organic" laundry detergent and make recommendations to customers about how to best use the product. They intentionally get grass stains on 30 white T-shirts in order to see how well the detergent will clean them. They want to try the detergent in cold water and in hot water on both the "regular" and "delicates" wash cycles. Design an appropriate experiment, indicating the number of factors, levels, and treatments. Explain the role of randomization in your experiment.

47. **Skydiving, anyone?** A humor piece published in the *British Medical Journal* ("Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized control trials," Gordon, Smith, and Pell, *BMJ*, 2003:327) notes that we can't tell for sure whether parachutes are safe and effective because there has never been a properly randomized, double-blind, placebo-controlled study of parachute effectiveness in skydiving. (Yes, this is the sort of thing statisticians find funny . . . .) Suppose you were designing such a study:
- a) What is the factor in this experiment?
- b) What experimental units would you propose?<sup>7</sup>
- c) What would serve as a placebo for this study?
- d) What would the treatments be?
- e) What would the response variable be?
- f) What sources of variability would you control?
- g) How would you randomize this "experiment"?
- h) How would you make the experiment double-blind?



## JUST CHECKING Answers

1. **a)** The factor was type of treatment for peptic ulcer.  
**b)** The response variable could be a measure of relief from gastric ulcer pain or an evaluation by a physician of the state of the disease.  
**c)** Treatments would be gastric freezing and some alternative control treatment.  
**d)** Treatments should be assigned randomly.  
**e)** No. The Web site reports "lack of effectiveness," indicating that no large differences in patient healing were noted.
2. **a)** Neither the patients who received the treatment nor the doctor who evaluated them afterward knew what treatment they had received.  
**b)** The placebo is needed to accomplish blinding. The best alternative would be using body-temperature liquid rather than the freezing liquid.  
**c)** The researchers should block the subjects by the length of time they had had the ulcer, then randomly assign subjects in each block to the freezing and placebo groups.

<sup>7</sup> Don't include your Statistics instructor!

## REVIEW OF PART III

## Gathering Data

## QUICK REVIEW

Before you can make a boxplot, calculate a mean, describe a distribution, or fit a line, you must have meaningful data to work with. Getting good data is essential to any investigation. No amount of clever analysis can make up for badly collected data. Here's a brief summary of the key concepts and skills:

- ▶ The way you gather data depends both on what you want to discover and on what is practical.
- ▶ To get some insight into what might happen in a real situation, model it with a **simulation** using random numbers.
- ▶ To answer questions about a target population, collect information from a sample with a **survey** or poll.
  - Choose the sample randomly. Random sampling designs include simple, stratified, systematic, cluster, and multistage.
  - A simple random sample draws without restriction from the entire target population.
  - When there are subgroups within the population that may respond differently, use a stratified sample.
  - Avoid bias, a systematic distortion of the results. Sample designs that allow undercoverage or response bias and designs such as voluntary response or convenience samples don't faithfully represent the population.
  - Samples will naturally vary one from another. This sample-to-sample variation is called sampling error. Each sample only approximates the target population.
- ▶ **Observational studies** collect information from a sample drawn from a target population.
  - Retrospective studies examine existing data. Prospective studies identify subjects in advance, then follow them to collect data as the data are created, perhaps over many years.
  - Observational studies can spot associations between variables but cannot establish cause and effect. It's impossible to eliminate the possibility of lurking or confounding variables.
- ▶ To see how different treatments influence a response variable, design an **experiment**.
  - Assign subjects to treatments randomly. If you don't assign treatments randomly, your experiment is not likely to yield valid results.
  - Control known sources of variation as much as possible. Reduce variation that cannot be controlled by using blocking, if possible.
  - Replicate the experiment, assigning several subjects to each treatment level.
  - If possible, replicate the entire experiment with an entirely different collection of subjects.
  - A well-designed experiment can provide evidence that changes in the factors cause changes in the response variable.

Now for more opportunities to review these concepts and skills . . .

## REVIEW EXERCISES

**1–18. What design?** Analyze the design of each research example reported. Is it a sample survey, an observational study, or an experiment? If a sample, what are the population, the parameter of interest, and the sampling procedure? If an observational study, was it retrospective or prospective? If an experiment, describe the factors, treatments, randomization, response variable, and any blocking, matching, or blinding that may be present. In each, what kind of conclusions can be reached?

1. Researchers identified 242 children in the Cleveland area who had been born prematurely (at about 29 weeks). They examined these children at age 8 and again at age 20, comparing them to another group of 233 children not born prematurely. Their report, published in the *New England Journal of Medicine*, said the “preemies” engaged in significantly less risky behavior than the others. Differences showed up in the use of alcohol and marijuana, conviction of crimes, and teenage pregnancy.
2. The journal *Circulation* reported that among 1900 people who had heart attacks, those who drank an average of 19 cups of tea a week were 44% more likely than non-drinkers to survive at least 3 years after the attack.
3. Researchers at the Purina Pet Institute studied Labrador retrievers for evidence of a relationship between diet and longevity. At 8 weeks of age, 2 puppies of the same sex and weight were randomly assigned to one of two groups—a total of 48 dogs in all. One group was allowed to eat all they wanted, while the other group was fed a diet about 25% lower in calories. The median lifespan of dogs fed the restricted diet was 22 months longer than that of other dogs. (*Science News* 161, no. 19)

4. The radioactive gas radon, found in some homes, poses a health risk to residents. To assess the level of contamination in their area, a county health department wants to test a few homes. If the risk seems high, they will publicize the results to emphasize the need for home testing. Officials plan to use the local property tax list to randomly choose 25 homes from various areas of the county.
5. Almost 90,000 women participated in a 16-year study of the role of the vitamin folate in preventing colon cancer. Some of the women had family histories of colon cancer in close relatives. In this at-risk group, the incidence of colon cancer was cut in half among those who maintained a high folate intake. No such difference was observed in those with no family-based risk. (*Science News*, Feb. 9, 2002)
6. In the journal *Science*, a research team reported that plants in southern England are flowering earlier in the spring. Records of the first flowering dates for 385 species over a period of 47 years indicate that flowering has advanced an average of 15 days per decade, an indication of climate warming, according to the authors.
7. Fireworks manufacturers face a dilemma. They must be sure that the rockets work properly, but test firing a rocket essentially destroys it. On the other hand, not testing the product leaves open the danger that they sell a bunch of duds, leading to unhappy customers and loss of future sales. The solution, of course, is to test a few of the rockets produced each day, assuming that if those tested work properly, the others are ready for sale.
8. Can makeup damage fetal development? Many cosmetics contain a class of chemicals called phthalates. Studies that exposed some laboratory animals to these chemicals found a heightened incidence of damage to male reproductive systems. Since traces of phthalates are found in the urine of women who use beauty products, there is growing concern that they may present a risk to male fetuses. (*Science News*, July 20, 2002)
9. Can long-term exposure to strong electromagnetic fields cause cancer? Researchers in Italy tracked down 13 years of medical records for people living near Vatican Radio's powerful broadcast antennas. A disproportionate share of the leukemia cases occurred among men and children who lived within 6 kilometers of the antennas. (*Science News*, July 20, 2002)
10. Some doctors have expressed concern that men who have vasectomies seemed more likely to develop prostate cancer. Medical researchers used a national cancer registry to identify 923 men who had had prostate cancer and 1224 men of similar ages who had not. Roughly one quarter of the men in each group had undergone a vasectomy, many more than 25 years before the study. The study's authors concluded that there is strong evidence that having the operation presents no long-term risk for developing prostate cancer. (*Science News*, July 20, 2002)
11. Researchers investigating appetite control as a means of losing weight found that female rats ate less and lost weight after injections of the hormone leptin, while male rats responded better to insulin. (*Science News*, July 20, 2002)
12. An artisan wants to create pottery that has the appearance of age. He prepares several samples of clay with four different glazes and test fires them in a kiln at three different temperature settings.
13. Tests of gene therapy on laboratory rats have raised hopes of stopping the degeneration of tissue that characterizes chronic heart failure. Researchers at the University of California, San Diego, used hamsters with cardiac disease, randomly assigning 30 to receive the gene therapy and leaving the other 28 untreated. Five weeks after treatment the gene therapy group's heart muscles stabilized, while those of the untreated hamsters continued to weaken. (*Science News*, July 27, 2002)
14. Researchers at the University of Bristol (England) investigated reasons why different species of birds begin to sing at different times in the morning. They captured and examined birds of 57 species at seven different sites. They measured the diameter of the birds' eyes and also recorded the time of day at which each species began to sing. These researchers reported a strong relationship between eye diameter and time of singing, saying that birds with bigger eyes tended to sing earlier. (*Science News*, 161, no. 16 [2002])
15. An orange-juice processing plant will accept a shipment of fruit only after several hundred oranges selected from various locations within the truck are carefully inspected. If too many show signs of unsuitability for juice (bruised, rotten, unripe, etc.), the whole truckload is rejected.
16. A soft-drink manufacturer must be sure the bottle caps on the soda are fully sealed and will not come off easily. Inspectors pull a few bottles off the production line at regular intervals and test the caps. If they detect any problems, they will stop the bottling process to adjust or repair the machine that caps the bottles.
17. Physically fit people seem less likely to die of cancer. A report in the May 2002 issue of *Medicine and Science in Sports and Exercise* followed 25,892 men aged 30 to 87 for 10 years. The most physically fit men had a 55% lower risk of death from cancer than the least fit group.
18. Does the use of computer software in Introductory Statistics classes lead to better understanding of the concepts? A professor teaching two sections of Statistics decides to investigate. She teaches both sections using the same lectures and assignments, but gives one class statistics software to help them with their homework. The classes take the same final exam, and graders do not know which students used computers during the semester. The professor is also concerned that students who have had calculus may perform differently from those who have not, so she plans to compare software vs. no-software scores separately for these two groups of students.
19. **Point spread.** When taking bets on sporting events, bookmakers often include a "point spread" that awards the weaker team extra points. In theory this makes the outcome of the bet a toss-up. Suppose a gambler places a \$10 bet and picks the winners of five games. If he's right about fewer than three of the games, he loses. If he gets three, four, or all five correct, he's paid \$10, \$20, and \$50, respectively. Estimate the amount such a bettor might expect to lose over many weeks of gambling.

20. **The lottery.** Many people spend a lot of money trying to win huge jackpots in state lotteries. Let's play a simplified version using only the numbers from 1 to 20. You bet on three numbers. The state picks five winning numbers. If your three are all among the winners, you are rich!
- Simulate repeated plays. How long did it take you to win?
  - In real lotteries, there are many more choices (often 54) and you must match all five winning numbers. Explain how these changes affect your chances of hitting the jackpot.
21. **Everyday randomness.** Aside from casinos, lotteries, and games, there are other situations you encounter in which something is described as "random" in some way. Give three different examples. Describe how randomness is (or is not) achieved in each.
22. **Cell phone risks.** Researchers at the Washington University School of Medicine randomly placed 480 rats into one of three chambers containing radio antennas. One group was exposed to digital cell phone radio waves, the second to analog cell phone waves, and the third group to no radio waves. Two years later the rats were examined for signs of brain tumors. In June 2002 the scientists said that differences among the three groups were not statistically significant.
- Is this a study or an experiment? Explain.
  - Explain in this context what "not statistically significant" means.
  - Comment on the fact that this research was funded by Motorola, a manufacturer of cell phones.
23. **Tips.** In restaurants, servers rely on tips as a major source of income. Does serving candy after the meal produce larger tips? To find out, two waiters determined randomly whether or not to give candy to 92 dining parties. They recorded the sizes of the tips and reported that guests getting candy tipped an average of 17.8% of the bill, compared with an average tip of only 15.1% from those who got no candy. ("Sweetening the Till: The Use of Candy to Increase Restaurant Tipping." *Journal of Applied Social Psychology* 32, no. 2 [2002]: 300–309)
- Was this an experiment or an observational study? Explain.
  - Is it reasonable to conclude that the candy caused guests to tip more? Explain.
  - The researchers said the difference was statistically significant. Explain in this context what that means.
24. **Tips, take 2.** In another experiment to see if getting candy after a meal would induce customers to leave a bigger tip, a waitress randomly decided what to do with 80 dining parties. Some parties received no candy, some just one piece, and some two pieces. Others initially got just one piece of candy, and then the waitress suggested that they take another piece. She recorded the tips received, finding that, in general, the more candy, the higher the tip, but the highest tips (23%) came from the parties who got one piece and then were offered more. ("Sweetening the Till: The Use of Candy to Increase Restaurant Tipping." *Journal of Applied Social Psychology* 32, no. 2 [2002]: 300–309)
- Diagram this experiment.
  - How many factors are there? How many levels?
  - How many treatments are there?
  - What is the response variable?
  - Did this experiment involve blinding? Double blinding?
  - In what way might the waitress, perhaps unintentionally, have biased the results?
25. **Cloning.** In September 1998, *USA Weekend* magazine asked, "Should humans be cloned?" Readers were invited to vote "Yes" or "No" by calling one of two different 900 numbers. Based on 38,023 responses, the magazine reported that "9 out of 10 readers oppose cloning."
- Explain why you think the conclusion is not justified. Describe the types of bias that may be present.
  - Reword the question in a way that you think might create a more positive response.
26. **Laundry.** An experiment to test a new laundry detergent, SparkleKleen, is being conducted by a consumer advocate group. They would like to compare its performance with that of a laboratory standard detergent they have used in previous experiments. They can stain 16 swatches of cloth with 2 tsp of a common staining compound and then use a well-calibrated optical scanner to detect the amount of the stain left after washing. To save time in the experiment, several suggestions have been made. Comment on the possible merits and drawbacks of each one.
- Since data for the laboratory standard detergent are already available from previous experiments, for this experiment wash all 16 swatches with SparkleKleen, and compare the results with the previous data.
  - Use both detergents with eight separate runs each, but to save time, use only a 10-second wash time with very hot water.
  - To ease bookkeeping, first run all of the standard detergent washes on eight swatches, then run all of the SparkleKleen washes on the other eight swatches.
  - Rather than run the experiment, use data from the company that produced SparkleKleen, and compare them with past data from the standard detergent.
27. **When to stop?** You play a game that involves rolling a die. You can roll as many times as you want, and your score is the total for all the rolls. But ... if you roll a 6 your score is 0 and your turn is over. What might be a good strategy for a game like this?
- One of your opponents decides to roll 4 times, then stop (hoping not to get the dreaded 6 before then). Use a simulation to estimate his average score.
  - Another opponent decides to roll until she gets at least 12 points, then stop. Use a simulation to estimate her average score.
  - Propose another strategy that you would use to play this game. Using your strategy, simulate several turns. Do you think you would beat the two opponents?
28. **Rivets.** A company that manufactures rivets believes the shear strength of the rivets they manufacture follows a Normal model with a mean breaking strength of 950 pounds and a standard deviation of 40 pounds.

- a) What percentage of rivets selected at random will break when tested under a 900-pound load?
- b) You're trying to improve the rivets and want to examine some that fail. Use a simulation to estimate how many rivets you might need to test in order to find three that fail at 900 pounds (or below).
29. **Homecoming.** A college Statistics class conducted a survey concerning community attitudes about the college's large homecoming celebration. That survey drew its sample in the following manner: Telephone numbers were generated at random by selecting one of the local telephone exchanges (first three digits) at random and then generating a random four-digit number to follow the exchange. If a person answered the phone and the call was to a residence, then that person was taken to be the subject for interview. (Undergraduate students and those under voting age were excluded, as was anyone who could not speak English.) Calls were placed until a sample of 200 eligible respondents had been reached.
- a) Did every telephone number that could occur in that community have an equal chance of being generated?
- b) Did this method of generating telephone numbers result in a simple random sample (SRS) of local residences? Explain.
- c) Did this method generate an SRS of local voters? Explain.
- d) Is this method unbiased in generating samples of households? Explain.
30. **Youthful appearance.** *Readers' Digest* reported results of several surveys that asked graduate students to examine photographs of men and women and try to guess their ages. Researchers compared these guesses with the number of times the people in the pictures reported having sexual intercourse. It turned out that those who had been more sexually active were judged as looking younger, and that the difference was described as "statistically significant." Psychologist David Weeks, who compiled the research, speculated that lovemaking boosts hormones that "reduce fatty tissue and increase lean muscle, giving a more youthful appearance."
- a) What does "statistically significant" mean in this context?
- b) Explain in statistical terms why you might be skeptical about Dr. Weeks's conclusion. Propose an alternative explanation for these results.
31. **Smoking and Alzheimer's.** Medical studies indicate that smokers are less likely to develop Alzheimer's disease than people who never smoked.
- a) Does this prove that smoking may offer some protection against Alzheimer's? Explain.
- b) Offer an alternative explanation for this association.
- c) How would you conduct a study to investigate this?
32. **Antacids.** A researcher wants to compare the performance of three types of antacid in volunteers suffering from acid reflux disease. Because men and women may react differently to this medication, the subjects are split into two groups, by sex. Subjects in each group are randomly assigned to take one of the antacids or to take a sugar pill made to look the same. The subjects will rate their level of discomfort 30 minutes after eating.
- a) What kind of design is this?
- b) The experiment uses volunteers rather than a random sample of all people suffering from acid reflux disease. Does this make the results invalid? Explain.
- c) How may the use of the placebo confound this experiment? Explain.
33. **Sex and violence.** Does the content of a television program affect viewers' memory of the products advertised in commercials? Design an experiment to compare the ability of viewers to recall brand names of items featured in commercials during programs with violent content, sexual content, or neutral content.
34. **Pubs.** In England, a Leeds University researcher said that the local watering hole's welcoming atmosphere helps men get rid of the stresses of modern life and is vital for their psychological well-being. Author of the report, Dr. Colin Gill, said rather than complain, women should encourage men to "pop out for a swift half." "Pub-time allows men to bond with friends and colleagues," he said. "Men need break-out time as much as women and are mentally healthier for it." Gill added that men might feel unfulfilled or empty if they had not been to the pub for a week. The report, commissioned by alcohol-free beer brand Kaliber, surveyed 900 men on their reasons for going to the pub. More than 40% said they went for the conversation, with relaxation and a friendly atmosphere being the other most common reasons. Only 1 in 10 listed alcohol as the overriding reason.
- Let's examine this news story from a statistical perspective.
- a) What are the W's: *Who, What, When, Where, Why?*
- b) What population does the researcher think the study applies to?
- c) What is the most important thing about the selection process that the article does *not* tell us?
- d) How do *you* think the 900 respondents were selected? (Name a method of drawing a sample that is likely to have been used.)
- e) Do you think the report that only 10% of respondents listed alcohol as an important reason for going to the pub might be a biased result? Why?
35. **Age and party.** The Gallup Poll conducted a representative telephone survey during the first quarter of 1999. Among its reported results was the following table concerning the preferred political party affiliation of respondents and their ages:

		Party			Total
		Republican	Democratic	Independent	
Age	18–29	241	351	409	1001
	30–49	299	330	370	999
	50–64	282	341	375	998
	65+	279	382	343	1004
Total		1101	1404	1497	4002

- a) What sampling strategy do you think the pollsters used? Explain.
- b) What percentage of the people surveyed were Democrats?
- c) Do you think this is a good estimate of the percentage of voters in the United States who are registered Democrats? Why or why not?
- d) In creating this sample design, what question do you think the pollsters were trying to answer?
36. **Bias?** Political analyst Michael Barone has written that “conservatives are more likely than others to refuse to respond to polls, particularly those polls taken by media outlets that conservatives consider biased” (*The Weekly Standard*, March 10, 1997). The Pew Research Foundation tested this assertion by asking the same questions in a national survey run by standard methods and in a more rigorous survey that was a true SRS with careful follow-up to encourage participation. The response rate in the “standard survey” was 42%. The response rate in the “rigorous survey” was 71%.
- a) What kind of bias does Barone claim may exist in polls?
- b) What is the population for these surveys?
- c) On the question of political position, the Pew researchers report the following table:

	Standard Survey	Rigorous Survey
Conservative	37%	35%
Moderate	40%	41%
Liberal	19%	20%

What makes you think these results are incomplete?

- d) The Pew researchers report that differences between opinions expressed on the two surveys were not statistically significant. Explain what “not statistically significant” means in this context.
37. **Save the grapes.** Vineyard owners have problems with birds that like to eat the ripening grapes. Some vineyards use scarecrows to try to keep birds away. Others use netting that covers the plants. Owners really would like to know if either method works and, if so, which one is better. One owner has offered to let you use his vineyard this year for an experiment. Propose a design. Carefully indicate how you would set up the experiment, specifying the factor(s) and response variable.
38. **Bats.** It’s generally believed that baseball players can hit the ball farther with aluminum bats than with the traditional wooden ones. Is that true? And, if so, how much farther? Players on your local high school baseball team have agreed to help you find out. Design an appropriate experiment.
39. **Knees.** Research reported in the spring of 2002 cast doubt on the effectiveness of arthroscopic knee surgery for patients with arthritis. Patients suffering from arthritis pain who volunteered to participate in the study were randomly divided into groups. One group received arthroscopic knee surgery. The other group underwent “placebo surgery” during which incisions were made in their knees, but no surgery was actually performed. Follow-up evaluations over a period of 2 years found that differences in the amount of pain relief experienced by the two groups were not statistically significant. (*NEJM* 347:81–88 July 11, 2002)
- a) Why did the researchers feel it was necessary to have some of the patients undergo “placebo surgery”?
- b) Because patients had to consent to participate in this experiment, the subjects were essentially self-selected—a kind of voluntary response group. Explain why that does not invalidate the findings of the experiment.
- c) What does “statistically significant” mean in this context?
40. **NBA draft lottery.** Professional basketball teams hold a “draft” each year in which they get to pick the best available college and high school players. In an effort to promote competition, teams with the worst records get to pick first, theoretically allowing them to add better players. To combat the fear that teams with no chance to make the playoffs might try to get better draft picks by intentionally losing late-season games, the NBA’s Board of Governors adopted a weighted lottery system in 1990. Under this system, the 11 teams that did not make the playoffs were eligible for the lottery. The NBA prepared 66 cards, each naming one of the teams. The team with the worst win-loss record was named on 11 of the cards, the second-worst team on 10 cards, and so on, with the team having the best record among the nonplayoff clubs getting only one chance at having the first pick. The cards were mixed, then drawn randomly to determine the order in which the teams could draft players. (Since 1995, 13 teams have been involved in the lottery, using a complicated system with 14 numbered Ping-Pong balls drawn in groups of four.) Suppose there are two exceptional players available in this year’s draft and your favorite team had the third-worst record. Use a simulation to find out how likely it is that your team gets to pick first or second. Describe your simulation carefully.
41. Security. There are 20 first-class passengers and 120 coach passengers scheduled on a flight. In addition to the usual security screening, 10% of the passengers will be subjected to a more complete search.
- a) Describe a sampling strategy to randomly select those to be searched.
- b) Here is the first-class passenger list and a set of random digits. Select two passengers to be searched, carefully demonstrating your process.
- 65436 71127 04879 41516 20451 02227 94769 23593
- |            |            |           |          |
|------------|------------|-----------|----------|
| Bergman    | Cox        | Fontana   | Perl     |
| Bowman     | DeLara     | Forester  | Rabkin   |
| Burkhauser | Delli-Bovi | Frongillo | Roufaiel |
| Castillo   | Dugan      | Furnas    | Swafford |
| Clancy     | Febo       | LePage    | Testut   |
- c) Explain how you would use a random number table to select the coach passengers to be searched.

42. **Profiling?** Among the 20 first-class passengers on the flight described in Exercise 41, there were four businessmen from the Middle East. Two of them were the two passengers selected to be searched. They complained of profiling, but the airline claims that the selection was random. What do you think? Support your conclusion with a simulation.
43. **Par 4.** In theory, a golfer playing a par-4 hole tees off, hitting the ball in the fairway, then hits an approach shot onto the green. The first putt (usually long) probably won't go in, but the second putt (usually much shorter) should. Sounds simple enough, but how many strokes might it really take? Use a simulation to estimate a pretty good golfer's score based on these assumptions:
- The tee shot hits the fairway 70% of the time.
  - A first approach shot lands on the green 80% of the time from the fairway, but only 40% of the time otherwise.
  - Subsequent approach shots land on the green 90% of the time.
  - The first putt goes in 20% of the time, and subsequent putts go in 90% of the time.
44. **The back nine.** Use simulations to estimate more golf scores, similar to the procedure in Exercise 43.
- a) On a par 3, the golfer hopes the tee shot lands on the green. Assume that the tee shot behaves like the first approach shot described in Exercise 43.
  - b) On a par 5, the second shot will reach the green 10% of the time and hit the fairway 60% of the time. If it does not hit the green, the golfer must play an approach shot as described in Exercise 43.
  - c) Create a list of assumptions that describe your golfing ability, and then simulate your score on a few holes. Explain your simulation clearly.





PART

IV

# Randomness and Probability

## Chapter 14

From Randomness to Probability

## Chapter 15

Probability Rules!

## Chapter 16

Random Variables

## Chapter 17

Probability Models

# From Randomness to Probability



Early humans saw a world filled with random events. To help them make sense of the chaos around them, they sought out seers, consulted oracles, and read tea leaves. As science developed, we learned to recognize some events as predictable. We can now forecast the change of seasons, tell when eclipses will occur precisely, and even make a reasonably good guess at how warm it will be tomorrow. But many other events are still essentially random. Will the stock market go up or down today? When will the next car pass this corner? And we now know from quantum mechanics that the universe is in some sense random at the most fundamental levels of subatomic particles.

But we have also learned to understand randomness. The surprising fact is that in the long run, even truly random phenomena settle down in a way that's consistent and predictable. It's this property of random phenomena that makes the next steps we're about to take in Statistics possible.

## Dealing with Random Phenomena

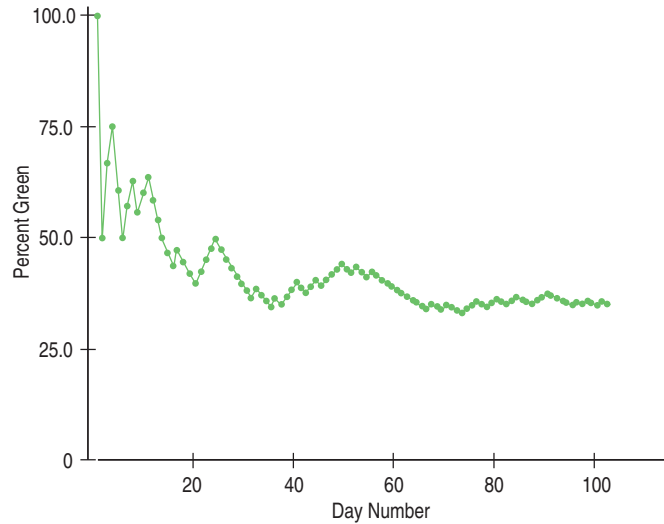
Every day you drive through the intersection at College and Main. Even though it may seem that the light is never green when you get there, you know this can't really be true. In fact, if you try really hard, you can recall just sailing through the green light once in a while.

What's random here? The light itself is governed by a timer. Its pattern isn't haphazard. In fact, the light may even be red at precisely the same times each day. It's the pattern of *your driving* that is random. No, we're certainly not insinuating that you can't keep the car on the road. At the precision level of the 30 seconds or so that the light spends being red or green, the time you arrive at the light *is random*. Even if you try to leave your house at exactly the same time every day, whether the light is red or green as *you* reach the intersection is a **random phenomenon**.<sup>1</sup>

<sup>1</sup> If you somehow managed to leave your house at *precisely* the same time every day and there was *no* variation in the time it took you to get to the light, then there wouldn't be any randomness, but that's not very realistic.

Is the color of the light completely unpredictable? When you stop to think about it, it's clear that you do expect some kind of *regularity* in your long-run experience. Some *fraction* of the time, the light will be green as you get to the intersection. How can you figure out what that fraction is?

You might record what happens at the intersection each day and graph the *accumulated percentage* of green lights like this:



**FIGURE 14.1**

The overall percentage of times the light is green settles down as you see more outcomes.

Day	Light	% Green
1	Green	100
2	Red	50
3	Green	66.7
4	Green	75
5	Red	60
6	Red	50
⋮	⋮	⋮

The first day you recorded the light, it was green. Then on the next five days, it was red, then green again, then green, red, and red. If you plot the percentage of green lights against days, the graph would start at 100% (because the first time, the light was green, so 1 out of 1, for 100%). Then the next day it was red, so the accumulated percentage dropped to 50% (1 out of 2). The third day it was green again (2 out of 3, or 67% green), then green (3 out of 4, or 75%), then red twice in a row (3 out of 5, for 60% green, and then 3 out of 6, for 50%), and so on. As you collect a new data value for each day, each new outcome becomes a smaller and smaller fraction of the accumulated experience, so, in the long run, the graph settles down. As it settles down, you can see that, in fact, the light is green about 35% of the time.

When talking about random phenomena such as this, it helps to define our terms. You aren't interested in the traffic light *all* the time. You pull up to the intersection only once a day, so you care about the color of the light only at these particular times.<sup>2</sup> In general, each occasion upon which we observe a random phenomenon is called a **trial**. At each trial, we note the value of the random phenomenon, and call that the trial's **outcome**. (If this language reminds you of Chapter 11, that's *not* unintentional.)

For the traffic light, there are really three possible outcomes: red, yellow, or green. Often we're more interested in a combination of outcomes rather than in the individual ones. When you see the light turn yellow, what do *you* do? If you race through the intersection, then you treat the yellow more like a green light. If you step on the brakes, you treat it more like a red light. Either way, you might want to group the yellow with one or the other. When we combine outcomes like that, the resulting combination is an **event**.<sup>3</sup> We sometimes talk about the collection of *all possible outcomes* and call that event the **sample space**.<sup>4</sup> We'll denote the sample

A phenomenon consists of trials. Each trial has an outcome. Outcomes combine to make events.

<sup>2</sup> Even though the randomness here comes from the uncertainty in our arrival time, we can think of the light itself as showing a color at random.

<sup>3</sup> Each individual outcome is also an event.

<sup>4</sup> Mathematicians like to use the term "space" as a fancy name for a set. Sort of like referring to that closet colleges call a dorm room as "living space." But remember that it's really just the set of all outcomes.

space  $S$ . (Some books are even fancier and use the Greek letter  $\Omega$ .) For the traffic light,  $S = \{\text{red, green, yellow}\}$ .

## The Law of Large Numbers



“For even the most stupid of men . . . is convinced that the more observations have been made, the less danger there is of wandering from one’s goal.”

—Jacob Bernoulli, 1713,  
discoverer of the LLN

### Empirical Probability

For any event  $A$ ,

$$P(A) = \frac{\text{\# times } A \text{ occurs}}{\text{total \# of trials}} \\ \text{in the long run.}$$

What’s the *probability* of a green light at College and Main? Based on the graph, it looks like the relative frequency of green lights settles down to about 35%, so saying that the probability is about 0.35 seems like a reasonable answer. But do random phenomena always behave well enough for this to make sense? Perhaps the relative frequency of an event can bounce back and forth between two values forever, never settling on just one number.

Fortunately, a principle called the **Law of Large Numbers** (LLN) gives us the guarantee we need. It simplifies things if we assume that the events are **independent**. Informally, this means that the outcome of one trial doesn’t affect the outcomes of the others. (We’ll see a formal definition of independent events in the next chapter.) The LLN says that as the number of independent trials increases, the long-run *relative frequency* of repeated events gets closer and closer to a single value.

Although the LLN wasn’t proven until the 18th century, everyone expects the kind of long-run regularity that the Law describes from everyday experience. In fact, the first person to prove the LLN, Jacob Bernoulli, thought it was pretty obvious, too, as his remark quoted in the margin shows.<sup>5</sup>

Because the LLN guarantees that relative frequencies settle down in the long run, we can now officially give a name to the value that they approach. We call it the **probability** of the event. If the relative frequency of green lights at that intersection settles down to 35% in the long run, we say that the probability of encountering a green light is 0.35, and we write  $P(\text{green}) = 0.35$ . Because this definition is based on repeatedly observing the event’s outcome, this definition of probability is often called **empirical probability**.

## The Nonexistent Law of Averages



Don’t let yourself think that there’s a Law of Averages that promises short-term compensation for recent deviations from expected behavior. A belief in such a “Law” can lead to money lost in gambling and to poor business decisions.

“Slump? I ain’t in no slump. I just ain’t hittin’.”

—Yogi Berra

Even though the LLN seems natural, it is often misunderstood because the idea of the *long run* is hard to grasp. Many people believe, for example, that an outcome of a random event that hasn’t occurred in many trials is “due” to occur. Many gamblers bet on numbers that haven’t been seen for a while, mistakenly believing that they’re likely to come up sooner. A common term for this is the “Law of Averages.” After all, we know that in the long run, the relative frequency will settle down to the probability of that outcome, so now we have some “catching up” to do, right?

Wrong. The Law of Large Numbers says nothing about short-run behavior. Relative frequencies even out *only in the long run*. And, according to the LLN, the long run is *really* long (*infinitely* long, in fact).

The so-called Law of Averages doesn’t exist at all. But you’ll hear people talk about it as if it does. Is a good hitter in baseball who has struck out the last six times *due* for a hit his next time up? If you’ve been doing particularly well in weekly quizzes in Statistics class, are you *due* for a bad grade? No. This isn’t the way random phenomena work. There is *no* Law of Averages for short runs.

The lesson of the LLN is that sequences of random events don’t compensate in the *short* run and don’t need to do so to get back to the right long-run probability.

<sup>5</sup>In case you were wondering, Jacob’s reputation was that he was every bit as nasty as this quotation suggests. He and his brother, who was also a mathematician, fought publicly over who had accomplished the most.

### The Law of Averages in Everyday Life

“Dear Abby: My husband and I just had our eighth child. Another girl, and I am really one disappointed woman. I suppose I should thank God she was healthy, but, Abby, this one was supposed to have been a boy. Even the doctor told me that the law of averages was in our favor 100 to one.” (Abigail Van Buren, 1974. Quoted in Karl Smith, *The Nature of Mathematics*. 6th ed. Pacific Grove, CA: Brooks/Cole, 1991, p. 589)

#### Ti-nspire

##### The Law of Large Numbers.

Watch the relative frequency of a random event approach the true probability in the long run.

If the probability of an outcome doesn't change and the events are independent, the probability of any outcome in another trial is *always* what it was, no matter what has happened in other trials.

**Coins, Keno, and the Law of Averages** You've just flipped a fair coin and seen six heads in a row. Does the coin “owe” you some tails? Suppose you spend that coin and your friend gets it in change. When she starts flipping the coin, should she expect a run of tails? Of course not. Each flip is a new event. The coin can't “remember” what it did in the past, so it can't “owe” any particular outcomes in the future.

Just to see how this works in practice, we ran a simulation of 100,000 flips of a fair coin. We collected 100,000 random numbers, letting the numbers 0 to 4 represent heads and the numbers 5 to 9 represent tails. In our 100,000 “flips,” there were 2981 streaks of at least 5 heads. The “Law of Averages” suggests that the next flip after a run of 5 heads should be tails more often to even things out. Actually, the next flip was heads more often than tails: 1550 times to 1431 times. That's 51.9% heads. You can perform a similar simulation easily on a computer. Try it!

Of course, sometimes an apparent drift from what we expect means that the probabilities are, in fact, *not* what we thought. If you get 10 heads in a row, maybe the coin has heads on both sides!

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80

Keno is a simple casino game in which numbers from 1 to 80 are chosen. The numbers, as in most lottery games, are supposed to be equally likely. Payoffs are made depending on how many of those numbers you match on your card. A group of graduate students from a Statistics department decided to take a field trip to Reno. They (*very discreetly*) wrote down the outcomes of the games for a couple of days, then drove back to test whether the numbers were, in fact, equally likely. It turned out that some numbers were *more likely* to come up than others. Rather than bet on the Law of Averages and put their

money on the numbers that were “due,” the students put their faith in the LLN—and all their (and their friends’) money on the numbers that had come up before. After they pocketed more than \$50,000, they were escorted off the premises and invited never to show their faces in that casino again.



## JUST CHECKING

1. One common proposal for beating the lottery is to note which numbers have come up lately, eliminate those from consideration, and bet on numbers that have not come up for a long time. Proponents of this method argue that in the long run, every number should be selected equally often, so those that haven't come up are due. Explain why this is faulty reasoning.

## Modeling Probability

### A S

**Activity: What Is Probability?** The best way to get a feel for probabilities is to experiment with them. We'll use this random-outcomes tool many more times.

Probability was first studied extensively by a group of French mathematicians who were interested in games of chance.<sup>6</sup> Rather than *experiment* with the games (and risk losing their money), they developed mathematical models of **theoretical probability**. To make things simple (as we usually do when we build models), they started by looking at games in which the different outcomes were equally likely. Fortunately, many games of chance are like that. Any of 52 cards is equally

<sup>6</sup> Ok, gambling.



### NOTATION ALERT:

We often use capital letters—and usually from the beginning of the alphabet—to denote events. We *always* use  $P$  to denote probability. So,

$$P(A) = 0.35$$

means “the probability of the event  $A$  is 0.35.”

When being formal, use decimals (or fractions) for the probability values, but sometimes, especially when talking more informally, it's easier to use percentages.

**A S**

**Activity: Multiple Discrete Outcomes.** The world isn't all heads or tails. Experiment with an event with 4 random alternative outcomes.

likely to be the next one dealt from a well-shuffled deck. Each face of a die is equally likely to land up (or at least it *should be*).

It's easy to find probabilities for events that are made up of several *equally likely* outcomes. We just count all the outcomes that the event contains. The probability of the event is the number of outcomes in the event divided by the total number of possible outcomes. We can write

$$P(A) = \frac{\# \text{ outcomes in } A}{\# \text{ of possible outcomes}}.$$

For example, the probability of drawing a face card (JQK) from a deck is

$$P(\text{face card}) = \frac{\# \text{ face cards}}{\# \text{ cards}} = \frac{12}{52} = \frac{3}{13}.$$

**Is that all there is to it?** Finding the probability of any event when the outcomes are equally likely is straightforward, but not necessarily easy. It gets hard when the number of outcomes in the event (and in the sample space) gets big. Think about flipping two coins. The sample space is  $S = \{HH, HT, TH, TT\}$  and each outcome is equally likely. So, what's the probability of getting *exactly* one head and one tail? Let's call that event  $A$ . Well, there are two outcomes in the event  $A = \{HT, TH\}$  out of the 4 possible equally likely ones in  $S$ , so  $P(A) = \frac{2}{4}$ , or  $\frac{1}{2}$ .

OK, now flip 100 coins. What's the probability of exactly 67 heads? Well, first, how many outcomes are in the sample space?  $S = \{HHHHHHHHHHH \dots H, HH \dots T, \dots\}$  Hmm. A lot. In fact, there are 1,267,650,600,228,229,401,496,703,205,376 different outcomes possible when flipping 100 coins. To answer the question, we'd still have to figure out how many ways there are to get 67 heads. That's coming in Chapter 17; stay tuned!

Don't get trapped into thinking that random events are always equally likely. The chance of winning a lottery—especially lotteries with very large payoffs—is small. Regardless, people continue to buy tickets. In an attempt to understand why, an interviewer asked someone who had just purchased a lottery ticket, “What do you think your chances are of winning the lottery?” The reply was, “Oh, about 50–50.” The shocked interviewer asked, “How do you get that?” to which the response was, “Well, the way I figure it, either I win or I don't!”

The moral of this story is that events are *not* always equally likely.

## Personal Probability

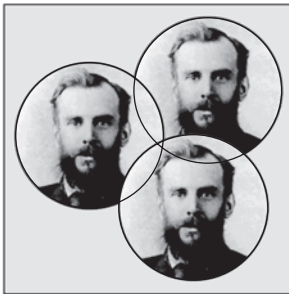
What's the probability that your grade in this Statistics course will be an  $A$ ? You may be able to come up with a number that seems reasonable. Of course, no matter how confident or depressed you feel about your chances for success, your probability should be between 0 and 1. How did you come up with this probability? Is it an empirical probability? Not unless you plan on taking the course over and over (and over . . .), calculating the proportion of times you get an  $A$ . And, unless you assume the outcomes are equally likely, it will be hard to find the theoretical probability. But people use probability in a third sense as well.

We use the language of probability in everyday speech to express a degree of uncertainty *without* basing it on long-run relative frequencies or mathematical models. Your personal assessment of your chances of getting an  $A$  expresses your

uncertainty about the outcome. That uncertainty may be based on how comfortable you're feeling in the course or on your midterm grade, but it can't be based on long-run behavior. We call this third kind of probability a subjective or **personal probability**.

Although personal probabilities may be based on experience, they're not based either on long-run relative frequencies or on equally likely events. So they don't display the kind of consistency that we'll need probabilities to have. For that reason, we'll stick to formally defined probabilities. You should be alert to the difference.

## The First Three Rules for Working with Probability



1. Make a picture.
2. Make a picture.
3. Make a picture.

We're dealing with probabilities now, not data, but the three rules don't change. The most common kind of picture to make is called a Venn diagram. We'll use Venn diagrams throughout the rest of this chapter. Even experienced statisticians make Venn diagrams to help them think about probabilities of compound and overlapping events. You should, too.

John Venn (1834–1923) created the Venn diagram. His book on probability, *The Logic of Chance*, was “strikingly original and considerably influenced the development of the theory of Statistics,” according to John Maynard Keynes, one of the luminaries of *Economics*.

## Formal Probability

### Surprising Probabilities

We've been careful to discuss probabilities only for situations in which the outcomes were finite, or even countably infinite. But if the outcomes can take on *any* numerical value at all (we say they are *continuous*), things can get surprising. For example, what is the probability that a randomly selected child will be *exactly* 3 feet tall? Well, if we mean 3.00000 . . . feet, the answer is zero. No randomly selected child—even one whose height would be recorded as 3 feet, will be *exactly* 3 feet tall (to an infinite number of decimal places). But, if you've grown taller than 3 feet, there must have been a time in your life when you actually *were* exactly 3 feet tall, even if only for a second. So this is an outcome with probability 0 that not only has happened—it has happened to *you*.

We've seen another example of this already in Chapter 6 when we worked with the Normal model. We said that the probability of any *specific* value—say,  $z = 0.5$ —is zero. The model gives a probability for any *interval* of values, such as  $0.49 < z < 0.51$ . The probability is smaller if we ask for  $0.499 < z < 0.501$ , and smaller still for  $0.49999999 < z < 0.50000001$ . Well, you get the idea. Continuous probabilities are useful for the mathematics behind much of what we'll do, but it's easier to deal with probabilities for countable outcomes.

For some people, the phrase “50/50” means something vague like “I don't know” or “whatever.” But when we discuss probabilities of outcomes, it takes on the precise meaning of *equally likely*. Speaking vaguely about probabilities will get us into trouble, so whenever we talk about probabilities, we'll need to be precise.<sup>7</sup> And to do that, we'll need to develop some formal rules<sup>8</sup> about how probability works.

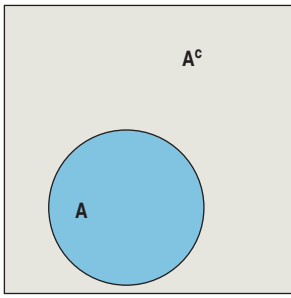
1. If the probability is 0, the event can't occur, and likewise if it has probability 1, it *always* occurs. Even if you think an event is very unlikely, its probability can't be negative, and even if you're sure it will happen, its probability can't be greater than 1. So we require that

**A probability is a number between 0 and 1.**

**For any event A,  $0 \leq P(A) \leq 1$ .**

<sup>7</sup> And to be precise, we will be talking only about sample spaces where we can enumerate all the outcomes. Mathematicians call this a countable number of outcomes.

<sup>8</sup> Actually, in mathematical terms, these are axioms—statements that we assume to be true of probability. We'll derive other rules from these in the next chapter.



The set **A** and its complement **A<sup>c</sup>**. Together, they make up the entire sample space **S**.

### NOTATION ALERT:

We write  $P(A \text{ or } B)$  as  $P(A \cup B)$ . The symbol  $\cup$  means “union,” representing the outcomes in event **A** or event **B** (or both). The symbol  $\cap$  means “intersection,” representing outcomes that are in both event **A** and event **B**. We write  $P(A \text{ and } B)$  as  $P(A \cap B)$ .

2. If a random phenomenon has only one possible outcome, it’s not very interesting (or very random). So we need to distribute the probabilities among all the outcomes a trial can have. How can we do that so that it makes sense? For example, consider what you’re doing as you read this book. The possible outcomes might be

- A:** You read to the end of this chapter before stopping.  
**B:** You finish this section but stop reading before the end of the chapter.  
**C:** You bail out before the end of this section.

When we assign probabilities to these outcomes, the first thing to be sure of is that we distribute all of the available probability. Something always occurs, so the probability of the entire sample space is 1.

Making this more formal gives the **Probability Assignment Rule**.

**The set of all possible outcomes of a trial must have probability 1.**

$$P(S) = 1$$

3. Suppose the probability that you get to class on time is 0.8. What’s the probability that you don’t get to class on time? Yes, it’s 0.2. The set of outcomes that are *not* in the event **A** is called the **complement of A**, and is denoted **A<sup>c</sup>**. This leads to the **Complement Rule**:

**The probability of an event occurring is 1 minus the probability that it doesn’t occur.**

$$P(A) = 1 - P(A^c)$$

### FOR EXAMPLE

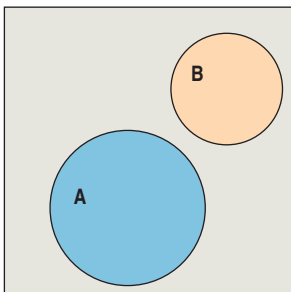
#### Applying the Complement Rule

**Recap:** We opened the chapter by looking at the traffic light at the corner of College and Main, observing that when we arrive at that intersection, the light is green about 35% of the time.

**Question:** If  $P(\text{green}) = 0.35$ , what’s the probability the light isn’t green when you get to College and Main?

$$\begin{aligned} \text{“Not green” is the complement of “green,” so } P(\text{not green}) &= 1 - P(\text{green}) \\ &= 1 - 0.35 = 0.65 \end{aligned}$$

There’s a 65% chance I won’t have a green light.



Two disjoint sets, **A** and **B**.

4. Suppose the probability that (**A**) a randomly selected student is a sophomore is 0.20, and the probability that (**B**) he or she is a junior is 0.30. What is the probability that the student is *either* a sophomore *or* a junior, written  $P(A \cup B)$ ? If you guessed 0.50, you’ve deduced the **Addition Rule**, which says that you can add the probabilities of events that are disjoint. To see whether two events are disjoint, we take them apart into their component outcomes and check whether they have any outcomes in common. **Disjoint (or mutually exclusive) events have no outcomes in common.** The **Addition Rule** states,

**For two disjoint events A and B, the probability that one or the other occurs is the sum of the probabilities of the two events.**

$$P(A \cup B) = P(A) + P(B), \text{ provided that A and B are disjoint.}$$



## FOR EXAMPLE

## Applying the Addition Rule

**Recap:** When you get to the light at College and Main, it's either red, green, or yellow. We know that  $P(\text{green}) = 0.35$ .

**Question:** Suppose we find out that  $P(\text{yellow})$  is about 0.04. What's the probability the light is red?

To find the probability that the light is green or yellow, I can use the Addition Rule because these are disjoint events: The light can't be both green and yellow at the same time.

$$P(\text{green} \cup \text{yellow}) = 0.35 + 0.04 = 0.39$$

Red is the only remaining alternative, and the probabilities must add up to 1, so

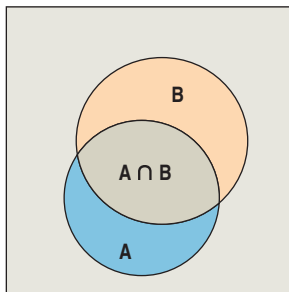
$$\begin{aligned} P(\text{red}) &= P(\text{not}(\text{green} \cup \text{yellow})) \\ &= 1 - P(\text{green} \cup \text{yellow}) \\ &= 1 - 0.39 = 0.61 \end{aligned}$$

"Baseball is 90% mental. The other half is physical."

—Yogi Berra

## AS

**Activity: Addition Rule for Disjoint Events.** Experiment with disjoint events to explore the Addition Rule.



Two sets **A** and **B** that are not disjoint. The event  $(A \cap B)$  is their intersection.

Because sample space outcomes are disjoint, we have an easy way to check whether the probabilities we've assigned to the possible outcomes are **legitimate**. The Probability Assignment Rule tells us that the sum of the probabilities of all possible outcomes must be exactly 1. No more, no less. For example, if we were told that the probabilities of selecting at random a freshman, sophomore, junior, or senior from all the undergraduates at a school were 0.25, 0.23, 0.22, and 0.20, respectively, we would know that something was wrong. These "probabilities" sum to only 0.90, so this is not a legitimate probability assignment. Either a value is wrong, or we just missed some possible outcomes, like "pre-freshman" or "postgraduate" categories that soak up the remaining 0.10. Similarly, a claim that the probabilities were 0.26, 0.27, 0.29, and 0.30 would be wrong because these "probabilities" sum to more than 1.

But be careful: The Addition Rule doesn't work for events that aren't disjoint. If the probability of owning an MP3 player is 0.50 and the probability of owning a computer is 0.90, the probability of owning either an MP3 player or a computer may be pretty high, but it is *not* 1.40! Why can't you add probabilities like this? Because these events are not disjoint. You *can* own both. In the next chapter, we'll see how to add probabilities for events like these, but we'll need another rule.

5. Suppose your job requires you to fly from Atlanta to Houston every Monday morning. The airline's Web site reports that this flight is on time 85% of the time. What's the chance that it will be on time two weeks in a row? That's the same as asking for the probability that your flight is on time this week *and* it's on time again next week. For independent events, the answer is very simple. Remember that independence means that the outcome of one event doesn't influence the outcome of the other. What happens with your flight this week doesn't influence whether it will be on time next week, so it's reasonable to assume that those events are independent. The **Multiplication Rule** says that for independent events, to find the probability that both events occur, we just multiply the probabilities together. Formally,

**For two independent events A and B, the probability that both A and B occur is the product of the probabilities of the two events.**

$$P(A \cap B) = P(A) \times P(B), \text{ provided that } A \text{ and } B \text{ are independent.}$$

AS

**Activity: Multiplication Rule for Independent Events.** Experiment with independent random events to explore the Multiplication Rule.

This rule can be extended to more than two independent events. What's the chance of your flight being on time for a month—four Mondays in a row? We can multiply the probabilities of it happening each week:

$$0.85 \times 0.85 \times 0.85 \times 0.85 = 0.522$$

or just over 50–50. Of course, to calculate this probability, we have used the assumption that the four events are independent.

Many Statistics methods require an **Independence Assumption**, but *assuming* independence doesn't make it true. Always *Think* about whether that assumption is reasonable before using the Multiplication Rule.

AS

**Activity: Probabilities of Compound Events.** The Random tool also lets you experiment with Compound random events to see if they are independent.

## FOR EXAMPLE

### Applying the Multiplication Rule (and others)

**Recap:** We've determined that the probability that we encounter a green light at the corner of College and Main is 0.35, a yellow light 0.04, and a red light 0.61. Let's think about your morning commute in the week ahead.

**Question:** What's the probability you find the light red both Monday and Tuesday?

Because the color of the light I see on Monday doesn't influence the color I'll see on Tuesday, these are independent events; I can use the Multiplication Rule:

$$\begin{aligned} P(\text{red Monday} \cap \text{red Tuesday}) &= P(\text{Red}) \times P(\text{red}) \\ &= (0.61)(0.61) \\ &= 0.3721 \end{aligned}$$

There's about a 37% chance I'll hit red lights both Monday and Tuesday mornings.

**Question:** What's the probability you don't encounter a red light until Wednesday?

For that to happen, I'd have to see green or yellow on Monday, green or yellow on Tuesday, and then red on Wednesday. I can simplify this by thinking of it as not red on Monday and Tuesday and then red on Wednesday.

$$\begin{aligned} P(\text{not red}) &= 1 - P(\text{red}) = 1 - 0.61 = 0.39, \text{ so} \\ P(\text{not red Monday} \cap \text{not red Tuesday} \cap \text{red Wednesday}) &= P(\text{not red}) \times P(\text{not red}) \times P(\text{red}) \\ &= (0.39)(0.39)(0.61) \\ &= 0.092781 \end{aligned}$$

There's about a 9% chance that this week I'll hit my first red light there on Wednesday morning.

**Question:** What's the probability that you'll have to stop *at least once* during the week?

Having to stop at least once means that I have to stop for the light either 1, 2, 3, 4, or 5 times next week. It's easier to think about the complement: never having to stop at a red light. Having to stop at least once means that I didn't make it through the week with no red lights.

$$\begin{aligned} P(\text{having to stop at the light at least once in 5 days}) &= 1 - P(\text{no red lights for 5 days in a row}) \\ &= 1 - P(\text{not red} \cap \text{not red} \cap \text{not red} \cap \text{not red} \cap \text{not red}) \\ &= 1 - (0.39)(0.39)(0.39)(0.39)(0.39) \\ &= 1 - 0.0090 \\ &= 0.991 \end{aligned}$$

There's over a 99% chance I'll hit at least one red light sometime this week.

Note that the phrase "at least" is often a tip-off to think about the complement. Something that happens *at least once* does happen. Happening at least once is the complement of not happening at all, and that's easier to find.

In informal English, you may see "some" used to mean "at least one." "What's the probability that some of the eggs in that carton are broken?" means at least one.



## JUST CHECKING

2. Opinion polling organizations contact their respondents by telephone. Random telephone numbers are generated, and interviewers try to contact those households. In the 1990s this method could reach about 69% of U.S. households. According to the Pew Research Center for the People and the Press, by 2003 the contact rate had risen to 76%. We can reasonably assume each household's response to be independent of the others. What's the probability that . . .
- a) the interviewer successfully contacts the next household on her list?
  - b) the interviewer successfully contacts both of the next two households on her list?
  - c) the interviewer's first successful contact is the third household on the list?
  - d) the interviewer makes at least one successful contact among the next five households on the list?

## STEP-BY-STEP EXAMPLE

### Probability



The five rules we've seen can be used in a number of different combinations to answer a surprising number of questions. Let's try one to see how we might go about it.

In 2001, Masterfoods, the manufacturers of M&M's<sup>®</sup> milk chocolate candies, decided to add another color to the standard color lineup of brown, yellow, red, orange, blue, and green. To decide which color to add, they surveyed people in nearly every country of the world and asked them to vote among purple, pink, and teal. The global winner was purple!

In the United States, 42% of those who voted said purple, 37% said teal, and only 19% said pink. But in Japan the percentages were 38% pink, 36% teal, and only 16% purple. Let's use Japan's percentages to ask some questions:

1. What's the probability that a Japanese M&M's survey respondent selected at random preferred either pink or teal?
2. If we pick two respondents at random, what's the probability that they both selected purple?
3. If we pick three respondents at random, what's the probability that *at least one* preferred purple?

### THINK

The probability of an event is its long-term relative frequency. It can be determined in several ways: by looking at many replications of an event, by deducing it from equally likely events, or by using some other information. Here, we are told the relative frequencies of the three responses.

Make sure the probabilities are legitimate. Here, they're not. Either there was a mistake, or the other voters must have chosen a color other than the three given. A check of the reports from other countries shows a similar deficit, so probably we're seeing those who had no preference or who wrote in another color.

The M&M's Web site reports the proportions of Japanese votes by color. These give the probability of selecting a voter who preferred each of the colors:

$$\begin{aligned} P(\text{pink}) &= 0.38 \\ P(\text{teal}) &= 0.36 \\ P(\text{purple}) &= 0.16 \end{aligned}$$

Each is between 0 and 1, but they don't all add up to 1. The remaining 10% of the voters must have not expressed a preference or written in another color. I'll put them together into "no preference" and add  $P(\text{no preference}) = 0.10$ .

With this addition, I have a legitimate assignment of probabilities.

<p><b>Question 1.</b> What's the probability that a Japanese M&amp;M's survey respondent selected at random preferred either pink or teal?</p>		
	<p><b>Plan</b> Decide which rules to use and check the conditions they require.</p>	<p>The events "Pink" and "Teal" are individual outcomes (a respondent can't choose both colors), so they are disjoint. I can apply the Addition Rule.</p>
	<p><b>Mechanics</b> Show your work.</p>	$P(\text{pink} \cup \text{teal}) = P(\text{pink}) + P(\text{teal})$ $= 0.38 + 0.36 = 0.74$
	<p><b>Conclusion</b> Interpret your results in the proper context.</p>	<p>The probability that the respondent said pink or teal is 0.74.</p>
<p><b>Question 2.</b> If we pick two respondents at random, what's the probability that they both said purple?</p>		
	<p><b>Plan</b> The word "both" suggests we want <math>P(\mathbf{A} \text{ and } \mathbf{B})</math>, which calls for the Multiplication Rule. Think about the assumption.</p>	<p>✓ <b>Independence Assumption:</b> It's unlikely that the choice made by one random respondent affected the choice of the other, so the events seem to be independent. I can use the Multiplication Rule.</p>
	<p><b>Mechanics</b> Show your work. For both respondents to pick purple, each one has to pick purple.</p>	$P(\text{both purple})$ $= P(\text{first respondent picks purple} \cap \text{second respondent picks purple})$ $= P(\text{first respondent picks purple}) \times P(\text{second respondent picks purple})$ $= 0.16 \times 0.16 = 0.0256$
	<p><b>Conclusion</b> Interpret your results in the proper context.</p>	<p>The probability that both respondents pick purple is 0.0256.</p>

**Question 3.** If we pick three respondents at random, what's the probability that at least one preferred purple?



**Plan** The phrase “at least . . .” often flags a question best answered by looking at the complement, and that's the best approach here. The complement of “At least one preferred purple” is “None of them preferred purple.”

Think about the assumption.

$$\begin{aligned} P(\text{at least one picked purple}) &= P(\{\text{none picked purple}\}^c) \\ &= 1 - P(\text{none picked purple}). \\ &= 1 - P(\text{not purple} \cap \text{not purple} \cap \text{not purple}). \end{aligned}$$

✓ **Independence Assumption:** These are independent events because they are choices by three random respondents. I can use the Multiplication Rule.



**Mechanics** First we find  $P(\text{not purple})$  with the Complement Rule.

Next we calculate  $P(\text{none picked purple})$  by using the Multiplication Rule.

Then we can use the Complement Rule to get the probability we want.

$$\begin{aligned} P(\text{not purple}) &= 1 - P(\text{purple}) \\ &= 1 - 0.16 = 0.84 \end{aligned}$$


$$\begin{aligned} P(\text{at least one picked purple}) &= 1 - P(\text{none picked purple}) \\ &= 1 - P(\text{not purple} \cap \text{not purple} \cap \text{not purple}) \\ &= 1 - (0.84)(0.84)(0.84) \\ &= 1 - 0.5927 \\ &= 0.4073 \end{aligned}$$

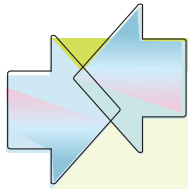


**Conclusion** Interpret your results in the proper context.

There's about a 40.7% chance that at least one of the respondents picked purple.

## WHAT CAN GO WRONG?

- ▶ **Beware of probabilities that don't add up to 1.** To be a legitimate probability assignment, the sum of the probabilities for all possible outcomes must total 1. If the sum is less than 1, you may need to add another category (“other”) and assign the remaining probability to that outcome. If the sum is more than 1, check that the outcomes are disjoint. If they're not, then you can't assign probabilities by just counting relative frequencies.
- ▶ **Don't add probabilities of events if they're not disjoint.** Events must be disjoint to use the Addition Rule. The probability of being under 80 *or* a female is not the probability of being under 80 *plus* the probability of being female. That sum may be more than 1.
- ▶ **Don't multiply probabilities of events if they're not independent.** The probability of selecting a student at random who is over 6'10" tall *and* on the basketball team is *not* the probability the student is over 6'10" tall *times* the probability he's on the basketball team. Knowing that the student is over 6'10" changes the probability of his being on the basketball team. You can't multiply these probabilities. The multiplication of probabilities of events that are not independent is one of the most common errors people make in dealing with probabilities.
- ▶ **Don't confuse disjoint and independent.** Disjoint events *can't* be independent. If  $A = \{\text{you get an A in this class}\}$  and  $B = \{\text{you get a B in this class}\}$ ,  $A$  and  $B$  are disjoint. Are they independent? If you find out that  $A$  is true, does that change the probability of  $B$ ? You bet it does! So they can't be independent. we'll return to this issue in the next chapter. 



## CONNECTIONS

We saw in the previous three chapters that randomness plays a critical role in gathering data. That fact alone makes it important that we understand how random events behave. The rules and concepts of probability give us a language to talk and think about random phenomena. From here on, randomness will be fundamental to how we think about data, and probabilities will show up in every chapter.

We began thinking about independence back in Chapter 3 when we looked at contingency tables and asked whether the distribution of one variable was the same for each category of another. Then, in Chapter 12, we saw that independence was fundamental to drawing a Simple Random Sample. For computing compound probabilities, we again ask about independence. And we'll continue to think about independence throughout the rest of the book.

Our interest in probability extends back to the start of the book. We've talked about "relative frequencies" often. But—let's be honest—that's just a casual term for probability. For example, you can now rephrase the 68–95–99.7 Rule to talk about the *probability* that a random value selected from a Normal model will fall within 1, 2, or 3 standard deviations of the mean.

Why not just say "probability" from the start? Well, we didn't need any of the formal rules of this chapter (or the next one), so there was no point to weighing down the discussion with those rules. And "relative frequency" is the right intuitive way to think about probability in this course, so you've been thinking right all along.

Keep it up.

## WHAT HAVE WE LEARNED?



We've learned that probability is based on long-run relative frequencies. We've thought about the Law of Large Numbers and noted that it speaks only of long-run behavior. Because the long run is a very long time, we need to be careful not to misinterpret the Law of Large Numbers. Even when we've observed a string of heads, we shouldn't expect extra tails in subsequent coin flips.

Also, we've learned some basic rules for combining probabilities of outcomes to find probabilities of more complex events. These include

- ▶ the Probability Assignment Rule,
- ▶ the Complement Rule,
- ▶ the Addition Rule for disjoint events, and
- ▶ the Multiplication Rule for independent events.

## Terms

Random phenomenon	324. A phenomenon is random if we know what outcomes could happen, but not which particular values will happen.
Trial	325. A single attempt or realization of a random phenomenon.
Outcome	325. The outcome of a trial is the value measured, observed, or reported for an individual instance of that trial.
Event	325. A collection of outcomes. Usually, we identify events so that we can attach probabilities to them. We denote events with bold capital letters such as <b>A</b> , <b>B</b> , or <b>C</b> .
Sample Space	325. The collection of all possible outcome values. The sample space has a probability of 1.
Law of Large Numbers	326. The Law of Large Numbers states that the long-run <i>relative frequency</i> of repeated independent events gets closer and closer to the <i>true</i> relative frequency as the number of trials increases.
Independence (informally)	326. Two events are <i>independent</i> if learning that one event occurs does not change the probability that the other event occurs.

Probability	326. The probability of an event is a number between 0 and 1 that reports the likelihood of that event's occurrence. We write $P(\mathbf{A})$ for the probability of the event $\mathbf{A}$ .
Empirical probability	326. When the probability comes from the long-run relative frequency of the event's occurrence, it is an <b>empirical probability</b> .
Theoretical probability	327. When the probability comes from a model (such as equally likely outcomes), it is called a <b>theoretical probability</b> .
Personal probability	328. When the probability is subjective and represents your personal degree of belief, it is called a <b>personal probability</b> .
The Probability Assignment Rule	330. The probability of the entire sample space must be 1. $P(\mathbf{S}) = 1$ .
Complement Rule	330. The probability of an event occurring is 1 minus the probability that it doesn't occur. $P(\mathbf{A}) = 1 - P(\mathbf{A}^c)$
Disjoint (Mutually exclusive)	330. Two events are disjoint if they share no outcomes in common. If $\mathbf{A}$ and $\mathbf{B}$ are disjoint, then knowing that $\mathbf{A}$ occurs tells us that $\mathbf{B}$ cannot occur. Disjoint events are also called "mutually exclusive."
Addition Rule	330. If $\mathbf{A}$ and $\mathbf{B}$ are disjoint events, then the probability of $\mathbf{A}$ or $\mathbf{B}$ is $P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B})$
Legitimate probability assignment	331. An assignment of probabilities to outcomes is legitimate if <ul style="list-style-type: none"> <li>▶ each probability is between 0 and 1 (inclusive).</li> <li>▶ the sum of the probabilities is 1.</li> </ul>
Multiplication Rule	331. If $\mathbf{A}$ and $\mathbf{B}$ are independent events, then the probability of $\mathbf{A}$ and $\mathbf{B}$ is $P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B})$
Independence Assumption	332. We often require events to be independent. (So you should think about whether this assumption is reasonable.)

## Skills

### THINK

- ▶ Understand that random phenomena are unpredictable in the short term but show long-run regularity.
- ▶ Be able to recognize random outcomes in a real-world situation.
- ▶ Know that the relative frequency of a random event settles down to a value called the (empirical) probability. Know that this is guaranteed for independent events by the Law of Large Numbers.
- ▶ Know the basic definitions and rules of probability.
- ▶ Recognize when events are disjoint and when events are independent. Understand the difference and that disjoint events cannot be independent.

### SHOW

- ▶ Be able to use the facts about probability to determine whether an assignment of probabilities is legitimate. Each probability must be a number between 0 and 1, and the sum of the probabilities assigned to all possible outcomes must be 1.
- ▶ Know how and when to apply the Addition Rule. Know that events must be disjoint for the Addition Rule to apply.
- ▶ Know how and when to apply the Multiplication Rule. Know that events must be independent for the Multiplication Rule to apply. Be able to use the Multiplication Rule to find probabilities for combinations of independent events.
- ▶ Know how to use the Complement Rule to make calculating probabilities simpler. Recognize that probabilities of "at least. . ." are likely to be simplified in this way.

### TELL

- ▶ Be able to use statements about probability in describing a random phenomenon. You will need this skill soon for making statements about statistical inference.
- ▶ Know and be able to use the terms "sample space", "disjoint events", and "independent events" correctly.

## EXERCISES

- Sample spaces.** For each of the following, list the sample space and tell whether you think the events are equally likely:
  - Toss 2 coins; record the order of heads and tails.
  - A family has 3 children; record the number of boys.
  - Flip a coin until you get a head or 3 consecutive tails; record each flip.
  - Roll two dice; record the larger number.

- Sample spaces.** For each of the following, list the sample space and tell whether you think the events are equally likely:
  - Roll two dice; record the sum of the numbers.
  - A family has 3 children; record each child's sex in order of birth.
  - Toss four coins; record the number of tails.
  - Toss a coin 10 times; record the length of the longest run of heads.

- Roulette.** A casino claims that its roulette wheel is truly random. What should that claim mean?

- Rain.** The weather reporter on TV makes predictions such as a 25% chance of rain. What do you think is the meaning of such a phrase?

- Winter.** Comment on the following quotation:

*"What I think is our best determination is it will be a colder than normal winter," said Pamela Naber Knox, a Wisconsin state climatologist. "I'm basing that on a couple of different things. First, in looking at the past few winters, there has been a lack of really cold weather. Even though we are not supposed to use the law of averages, we are due." (Associated Press, fall 1992, quoted by Schaeffer et al.)*

- Snow.** After an unusually dry autumn, a radio announcer is heard to say, "Watch out! We'll pay for these sunny days later on this winter." Explain what he's trying to say, and comment on the validity of his reasoning.

- Cold streak.** A batter who had failed to get a hit in seven consecutive times at bat then hits a game-winning home run. When talking to reporters afterward, he says he was very confident that last time at bat because he knew he was "due for a hit." Comment on his reasoning.

- Crash.** Commercial airplanes have an excellent safety record. Nevertheless, there are crashes occasionally, with the loss of many lives. In the weeks following a crash, airlines often report a drop in the number of passengers, probably because people are afraid to risk flying.

- A travel agent suggests that since the law of averages makes it highly unlikely to have two plane crashes within a few weeks of each other, flying soon after a crash is the safest time. What do you think?
- If the airline industry proudly announces that it has set a new record for the longest period of safe flights, would you be reluctant to fly? Are the airlines due to have a crash?

- Fire insurance.** Insurance companies collect annual payments from homeowners in exchange for paying to rebuild houses that burn down.

- Why should you be reluctant to accept a \$300 payment from your neighbor to replace his house should it burn down during the coming year?
- Why can the insurance company make that offer?

- Jackpot.** On January 20, 2000, the International Gaming Technology company issued a press release:

*(LAS VEGAS, Nev.)—Cynthia Jay was smiling ear to ear as she walked into the news conference at The Desert Inn Resort in Las Vegas today, and well she should. Last night, the 37-year-old cocktail waitress won the world's largest slot jackpot—\$34,959,458—on a Megabucks machine. She said she had played \$27 in the machine when the jackpot hit. Nevada Megabucks has produced 49 major winners in its 14-year history. The top jackpot builds from a base amount of \$7 million and can be won with a 3-coin (\$3) bet.*

- How can the Desert Inn afford to give away millions of dollars on a \$3 bet?
- Why did the company issue a press release? Wouldn't most businesses want to keep such a huge loss quiet?

- Spinner.** The plastic arrow on a spinner for a child's game stops rotating to point at a color that will determine what happens next. Which of the following probability assignments are possible?

	Probabilities of . . .			
	Red	Yellow	Green	Blue
a)	0.25	0.25	0.25	0.25
b)	0.10	0.20	0.30	0.40
c)	0.20	0.30	0.40	0.50
d)	0	0	1.00	0
e)	0.10	0.20	1.20	-1.50

- Scratch off.** Many stores run "secret sales": Shoppers receive cards that determine how large a discount they get, but the percentage is revealed by scratching off that black stuff (what is that?) only after the purchase has been totaled at the cash register. The store is required to reveal (in the fine print) the distribution of discounts available. Which of these probability assignments are legitimate?

	Probabilities of . . .			
	10% off	20% off	30% off	50% off
a)	0.20	0.20	0.20	0.20
b)	0.50	0.30	0.20	0.10
c)	0.80	0.10	0.05	0.05
d)	0.75	0.25	0.25	-0.25
e)	1.00	0	0	0



13. **Vehicles.** Suppose that 46% of families living in a certain county own a car and 18% own an SUV. The Addition Rule might suggest, then, that 64% of families own either a car or an SUV. What's wrong with that reasoning?
14. **Homes.** Funding for many schools comes from taxes based on assessed values of local properties. People's homes are assessed higher if they have extra features such as garages and swimming pools. Assessment records in a certain school district indicate that 37% of the homes have garages and 3% have swimming pools. The Addition Rule might suggest, then, that 40% of residences have a garage or a pool. What's wrong with that reasoning?
15. **Speeders.** Traffic checks on a certain section of highway suggest that 60% of drivers are speeding there. Since  $0.6 \times 0.6 = 0.36$ , the Multiplication Rule might suggest that there's a 36% chance that two vehicles in a row are both speeding. What's wrong with that reasoning?
16. **Lefties.** Although it's hard to be definitive in classifying people as right- or left-handed, some studies suggest that about 14% of people are left-handed. Since  $0.14 \times 0.14 = 0.0196$ , the Multiplication Rule might suggest that there's about a 2% chance that a brother and a sister are both lefties. What's wrong with that reasoning?
17. **College admissions.** For high school students graduating in 2007, college admissions to the nation's most selective schools were the most competitive in memory. (*The New York Times*, "A Great Year for Ivy League Schools, but Not So Good for Applicants to Them," April 4, 2007). Harvard accepted about 9% of its applicants, Stanford 10%, and Penn 16%. Jorge has applied to all three. Assuming that he's a typical applicant, he figures that his chances of getting into both Harvard and Stanford must be about 0.9%.
- How has he arrived at this conclusion?
  - What additional assumption is he making?
  - Do you agree with his conclusion?
18. **College admissions II.** In Exercise 17, we saw that in 2007 Harvard accepted about 9% of its applicants, Stanford 10%, and Penn 16%. Jorge has applied to all three. He figures that his chances of getting into at least one of the three must be about 35%.
- How has he arrived at this conclusion?
  - What assumption is he making?
  - Do you agree with his conclusion?
19. **Car repairs.** A consumer organization estimates that over a 1-year period 17% of cars will need to be repaired once, 7% will need repairs twice, and 4% will require three or more repairs. What is the probability that a car chosen at random will need
- no repairs?
  - no more than one repair?
  - some repairs?
20. **Stats projects.** In a large Introductory Statistics lecture hall, the professor reports that 55% of the students enrolled have never taken a Calculus course, 32% have taken only one semester of Calculus, and the rest have taken two or more semesters of Calculus. The professor randomly assigns students to groups of three to work on a project for the course. What is the probability that the first groupmate you meet has studied
- two or more semesters of Calculus?
  - some Calculus?
  - no more than one semester of Calculus?
21. **More repairs.** Consider again the auto repair rates described in Exercise 19. If you own two cars, what is the probability that
- neither will need repair?
  - both will need repair?
  - at least one car will need repair?
22. **Another project.** You are assigned to be part of a group of three students from the Intro Stats class described in Exercise 20. What is the probability that of your other two groupmates,
- neither has studied Calculus?
  - both have studied at least one semester of Calculus?
  - at least one has had more than one semester of Calculus?
23. **Repairs, again.** You used the Multiplication Rule to calculate repair probabilities for your cars in Exercise 21.
- What must be true about your cars in order to make that approach valid?
  - Do you think this assumption is reasonable? Explain.
24. **Final project.** You used the Multiplication Rule to calculate probabilities about the Calculus background of your Statistics groupmates in Exercise 22.
- What must be true about the groups in order to make that approach valid?
  - Do you think this assumption is reasonable? Explain.
25. **Energy 2007.** A Gallup poll in March 2007 asked 1005 U.S. adults whether increasing domestic energy production or protecting the environment should be given a higher priority. Here are the results:

Response	Number
Increase production	342
Protect environment	583
Equally important	30
No opinion	50
<b>Total</b>	<b>1005</b>

If we select a person at random from this sample of 1005 adults,

- what is the probability that the person responded "Increase production"?
- what is the probability that the person responded "Equally important" or had no opinion?

26. **Failing fathers?** A Pew Research poll in 2007 asked 2020 U.S. adults whether fathers today were doing as good a job of fathering as fathers of 20–30 years ago. Here's how they responded:

Response	Number
Better	424
Same	566
Worse	950
No Opinion	80
<b>Total</b>	<b>2020</b>

If we select a respondent at random from this sample of 2020 adults,

- what is the probability that the selected person responded "Worse"?
  - what is the probability that the person responded the "Same" or "Better"?
27. **More energy.** Exercise 25 shows the results of a Gallup Poll about energy. Suppose we select three people at random from this sample.
- What is the probability that all three responded "Protect the environment"?
  - What is the probability that none responded "Equally important"?
  - What assumption did you make in computing these probabilities?
  - Explain why you think that assumption is reasonable.
28. **Fathers revisited.** Consider again the results of the poll about fathering discussed in Exercise 26. If we select two people at random from this sample,
- what is the probability that both think fathers are better today?
  - what is the probability that neither thinks fathers are better today?
  - what is the probability that one person thinks fathers are better today and the other doesn't?
  - What assumption did you make in computing these probabilities?
  - Explain why you think that assumption is reasonable.
29. **Polling.** As mentioned in the chapter, opinion-polling organizations contact their respondents by sampling random telephone numbers. Although interviewers now can reach about 76% of U.S. households, the percentage of those contacted who agree to cooperate with the survey has fallen from 58% in 1997 to only 38% in 2003 (Pew Research Center for the People and the Press). Each household, of course, is independent of the others.
- What is the probability that the next household on the list will be contacted but will refuse to cooperate?
  - What is the probability (in 2003) of failing to contact a household or of contacting the household but not getting them to agree to the interview?
  - Show another way to calculate the probability in part b.
30. **Polling, part II.** According to Pew Research, the contact rate (probability of contacting a selected household) was 69% in 1997 and 76% in 2003. However, the cooperation rate (probability of someone at the contacted household agreeing to be interviewed) was 58% in 1997 and dropped to 38% in 2003.
- What is the probability (in 2003) of obtaining an interview with the next household on the sample list? (To obtain an interview, an interviewer must both contact the household and then get agreement for the interview.)
  - Was it more likely to obtain an interview from a randomly selected household in 1997 or in 2003?
31. **M&M's.** The Masterfoods company says that before the introduction of purple, yellow candies made up 20% of their plain M&M's, red another 20%, and orange, blue, and green each made up 10%. The rest were brown.
- If you pick an M&M at random, what is the probability that
    - it is brown?
    - it is yellow or orange?
    - it is not green?
    - it is striped?
  - If you pick three M&M's in a row, what is the probability that
    - they are all brown?
    - the third one is the first one that's red?
    - none are yellow?
    - at least one is green?
32. **Blood.** The American Red Cross says that about 45% of the U.S. population has Type O blood, 40% Type A, 11% Type B, and the rest Type AB.
- Someone volunteers to give blood. What is the probability that this donor
    - has Type AB blood?
    - has Type A or Type B?
    - is not Type O?
  - Among four potential donors, what is the probability that
    - all are Type O?
    - no one is Type AB?
    - they are not all Type A?
    - at least one person is Type B?
33. **Disjoint or independent?** In Exercise 31 you calculated probabilities of getting various M&M's. Some of your answers depended on the assumption that the outcomes described were *disjoint*; that is, they could not both happen at the same time. Other answers depended on the assumption that the events were *independent*; that is, the occurrence of one of them doesn't affect the probability of the other. Do you understand the difference between disjoint and independent?
- If you draw one M&M, are the events of getting a red one and getting an orange one disjoint, independent, or neither?
  - If you draw two M&M's one after the other, are the events of getting a red on the first and a red on the second disjoint, independent, or neither?
  - Can disjoint events ever be independent? Explain.
34. **Disjoint or independent?** In Exercise 32 you calculated probabilities involving various blood types. Some of your answers depended on the assumption that the outcomes described were *disjoint*; that is, they could not both happen at the same time. Other answers depended on the assumption that the events were *independent*; that is, the occurrence of one of them doesn't affect the probability of

the other. Do you understand the difference between disjoint and independent?

- a) If you examine one person, are the events that the person is Type A and that the person is Type B disjoint, independent, or neither?
- b) If you examine two people, are the events that the first is Type A and the second Type B disjoint, independent, or neither?
- c) Can disjoint events ever be independent? Explain.
35. **Dice.** You roll a fair die three times. What is the probability that
- a) you roll all 6's?
- b) you roll all odd numbers?
- c) none of your rolls gets a number divisible by 3?
- d) you roll at least one 5?
- e) the numbers you roll are not all 5's?
36. **Slot machine.** A slot machine has three wheels that spin independently. Each has 10 equally likely symbols: 4 bars, 3 lemons, 2 cherries, and a bell. If you play, what is the probability that
- a) you get 3 lemons?
- b) you get no fruit symbols?
- c) you get 3 bells (the jackpot)?
- d) you get no bells?
- e) you get at least one bar (an automatic loser)?
37. **Champion bowler.** A certain bowler can bowl a strike 70% of the time. What's the probability that she
- a) goes three consecutive frames without a strike?
- b) makes her first strike in the third frame?
- c) has at least one strike in the first three frames?
- d) bowls a perfect game (12 consecutive strikes)?
38. **The train.** To get to work, a commuter must cross train tracks. The time the train arrives varies slightly from day to day, but the commuter estimates he'll get stopped on about 15% of work days. During a certain 5-day work week, what is the probability that he
- a) gets stopped on Monday and again on Tuesday?
- b) gets stopped for the first time on Thursday?
- c) gets stopped every day?
- d) gets stopped at least once during the week?
39. **Voters.** Suppose that in your city 37% of the voters are registered as Democrats, 29% as Republicans, and 11% as members of other parties (Liberal, Right to Life, Green, etc.). Voters not aligned with any official party are termed "Independent." You are conducting a poll by calling registered voters at random. In your first three calls, what is the probability you talk to
- a) all Republicans?
- b) no Democrats?
- c) at least one Independent?
40. **Religion.** Census reports for a city indicate that 62% of residents classify themselves as Christian, 12% as Jewish, and 16% as members of other religions (Muslims, Buddhists, etc.). The remaining residents classify themselves as nonreligious. A polling organization seeking information about public opinions wants to be sure to talk with people holding a variety of religious views, and makes random phone calls. Among the first four people they call, what is the probability they reach
- a) all Christians?
- b) no Jews?
- c) at least one person who is nonreligious?
41. **Tires.** You bought a new set of four tires from a manufacturer who just announced a recall because 2% of those tires are defective. What is the probability that at least one of yours is defective?
42. **Pepsi.** For a sales promotion, the manufacturer places winning symbols under the caps of 10% of all Pepsi bottles. You buy a six-pack. What is the probability that you win something?
43. **9/11?** On September 11, 2002, the first anniversary of the terrorist attack on the World Trade Center, the New York State Lottery's daily number came up 9-1-1. An interesting coincidence or a cosmic sign?



- a) What is the probability that the winning three numbers match the date on any given day?
- b) What is the probability that a whole year passes without this happening?
- c) What is the probability that the date and winning lottery number match at least once during any year?
- d) If every one of the 50 states has a three-digit lottery, what is the probability that at least one of them will come up 9-1-1 on September 11?
44. **Red cards.** You shuffle a deck of cards and then start turning them over one at a time. The first one is red. So is the second. And the third. In fact, you are surprised to get 10 red cards in a row. You start thinking, "The next one is due to be black!"
- a) Are you correct in thinking that there's a higher probability that the next card will be black than red? Explain.
- b) Is this an example of the Law of Large Numbers? Explain.



### JUST CHECKING Answers

1. The LLN works only in the long run, not in the short run. The random methods for selecting lottery numbers have no memory of previous picks, so there is no change in the probability that a certain number will come up.
2. a) 0.76
- b)  $0.76(0.76) = 0.5776$
- c)  $(1 - 0.76)^2(0.76) = 0.043776$
- d)  $1 - (1 - 0.76)^5 = 0.9992$

# Probability Rules!



**P**ull a bill from your wallet or pocket without looking at it. An outcome of this trial is the bill you select. The sample space is all the bills in circulation:  $S = \{\$1 \text{ bill}, \$2 \text{ bill}, \$5 \text{ bill}, \$10 \text{ bill}, \$20 \text{ bill}, \$50 \text{ bill}, \$100 \text{ bill}\}$ .<sup>1</sup> These are *all* the possible outcomes. (In spite of what you may have seen in bank robbery movies, there are no \$500 or \$1000 bills.)

We can combine the outcomes in different ways to make many different events. For example, the event  $A = \{\$1, \$5, \$10\}$  represents selecting a \$1, \$5, or \$10 bill. The event  $B = \{\text{a bill that does not have a president on it}\}$  is the collection of outcomes (Don't look! Can you name them?):  $\{\$10 \text{ (Hamilton)}, \$100 \text{ (Franklin)}\}$ . The event  $C = \{\text{enough money to pay for a \$12 meal with one bill}\}$  is the set of outcomes  $\{\$20, \$50, \$100\}$ .

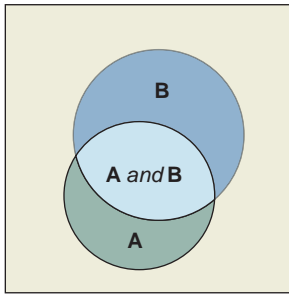
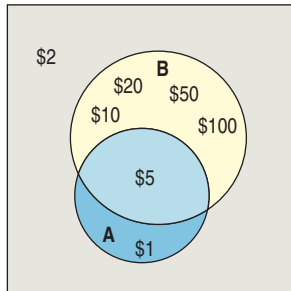
Notice that these outcomes are not equally likely. You'd no doubt be more surprised (and pleased) to pull out a \$100 bill than a \$1 bill—it's not very likely, though. You probably carry many more \$1 than \$100 bills, but without information about the probability of each outcome, we can't calculate the probability of an event.

The probability of the event  $C$  (getting a bill worth more than \$12) is *not*  $3/7$ . There are 7 possible outcomes, and 3 of them exceed \$12, but they are not *equally likely*. (Remember the probability that your lottery ticket will win rather than lose still isn't  $1/2$ .)

## The General Addition Rule

Now look at the bill in your hand. There are images of famous buildings in the center of the backs of all but two bills in circulation. The \$1 bill has the word ONE in the center, and the \$2 bill shows the signing of the Declaration of Independence.

<sup>1</sup> Well, technically, the sample space is all the bills in your pocket. You may be quite sure there isn't a \$100 bill in there, but *we* don't know that, so humor us that it's at least *possible* that any legal bill could be there.

Events **A** and **B** and their intersection.Denominations of bills that are odd (**A**) or that have a building on the reverse side (**B**). The two sets both include the \$5 bill, and both exclude the \$2 bill.

What's the probability of randomly selecting  $A = \{\text{a bill with an odd-numbered value}\}$  or  $B = \{\text{a bill with a building on the reverse}\}$ ? We know  $A = \{\$1, \$5\}$  and  $B = \{\$5, \$10, \$20, \$50, \$100\}$ . But  $P(A \text{ or } B)$  is not simply the sum  $P(A) + P(B)$ , because the events  $A$  and  $B$  are not disjoint. The \$5 bill is in both sets. So what can we do? We'll need a new probability rule.

As the diagrams show, we can't use the Addition Rule and add the two probabilities because the events are not disjoint; they overlap. There's an outcome (the \$5 bill) in the *intersection* of  $A$  and  $B$ . The Venn diagram represents the sample space. Notice that the \$2 bill has neither a building nor an odd denomination, so it sits outside both circles.

The \$5 bill plays a crucial role here because it is both odd *and* has a building on the reverse. It's in both  $A$  and  $B$ , which places it in the *intersection* of the two circles. The reason we can't simply add the probabilities of  $A$  and  $B$  is that we'd count the \$5 bill twice.

If we did add the two probabilities, we could compensate by *subtracting* out the probability of that \$5 bill. So,

$$\begin{aligned} P(\text{odd number value or building}) &= P(\text{odd number value}) + P(\text{building}) - P(\text{odd number value and building}) \\ &= P(\$1, \$5) + P(\$5, \$10, \$20, \$50, \$100) - P(\$5). \end{aligned}$$

This method works in general. We add the probabilities of two events and then subtract out the probability of their intersection. This approach gives us the **General Addition Rule**, which does not require disjoint events:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

**FOR EXAMPLE****Using the General Addition Rule**

A survey of college students found that 56% live in a campus residence hall, 62% participate in a campus meal program, and 42% do both.

**Question:** What's the probability that a randomly selected student either lives or eats on campus?

Let  $L = \{\text{student lives on campus}\}$  and  $M = \{\text{student has a campus meal plan}\}$ .

$$\begin{aligned} P(\text{a student either lives or eats on campus}) &= P(L \cup M) \\ &= P(L) + P(M) - P(L \cap M) \\ &= 0.56 + 0.62 - 0.42 \\ &= 0.76 \end{aligned}$$

There's a 76% chance that a randomly selected college student either lives or eats on campus.

**Would you like dessert or coffee?** Natural language can be ambiguous. In this question, is the answer one of the two alternatives, or simply "yes"? Must you decide between them, or may you have both? That kind of ambiguity can confuse our probabilities.

Suppose we had been asked a different question: What is the probability that the bill we draw has *either* an odd value *or* a building but *not both*? Which bills are we talking about now? The set we're interested in would be  $\{\$1, \$10, \$20, \$50, \$100\}$ . We don't include the \$5 bill in the set because it has both characteristics.

Why isn't this the same answer as before? The problem is that when we say the word "or," we usually mean *either one or both*. We don't usually mean the *exclusive*

version of “or” as in, “Would you like the steak *or* the vegetarian entrée?” Ordinarily when we ask for the probability that **A** or **B** occurs, we mean **A** or **B** or both. And we know *that* probability is  $P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \text{ and } \mathbf{B})$ . The General Addition Rule subtracts the probability of the outcomes in **A** and **B** because we’ve counted those outcomes *twice*. But they’re still there.

If we really mean **A** or **B** but NOT both, we have to get rid of the outcomes in **{A and B}**. So  $P(\mathbf{A} \text{ or } \mathbf{B} \text{ but not both}) = P(\mathbf{A} \cup \mathbf{B}) - P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - 2 \times P(\mathbf{A} \cap \mathbf{B})$ . Now we’ve subtracted  $P(\mathbf{A} \cap \mathbf{B})$  twice—once because we don’t want to double-count these events and a second time because we really didn’t want to count them at all.

Confused? *Make a picture*. It’s almost always easier to think about such situations by looking at a Venn diagram.

## FOR EXAMPLE

### Using Venn diagrams

**Recap:** We return to our survey of college students: 56% live on campus, 62% have a campus meal program, and 42% do both.

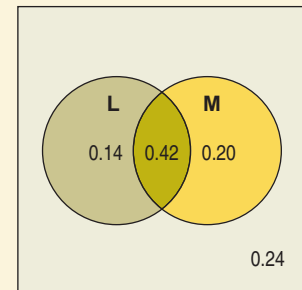
**Questions:** Based on a Venn diagram, what is the probability that a randomly selected student

- lives off campus and doesn’t have a meal program?
- lives in a residence hall but doesn’t have a meal program?

Let  $\mathbf{L} = \{\text{student lives on campus}\}$  and  $\mathbf{M} = \{\text{student has a campus meal plan}\}$ . In the Venn diagram, the intersection of the circles is  $P(\mathbf{L} \cap \mathbf{M}) = 0.42$ . Since  $P(\mathbf{L}) = 0.56$ ,  $P(\mathbf{L} \cap \mathbf{M}^c) = 0.56 - 0.42 = 0.14$ . Also,  $P(\mathbf{L}^c \cap \mathbf{M}) = 0.62 - 0.42 = 0.20$ . Now,  $0.14 + 0.42 + 0.20 = 0.76$ , leaving  $1 - 0.76 = 0.24$  for the region outside both circles.

Now . . .  $P(\text{off campus and no meal program}) = P(\mathbf{L}^c \cap \mathbf{M}^c) = 0.24$

$P(\text{on campus and no meal program}) = P(\mathbf{L} \cap \mathbf{M}^c) = 0.14$



## JUST CHECKING

- Back in Chapter 1 we suggested that you sample some pages of this book at random to see whether they held a graph or other data display. We actually did just that. We drew a representative sample and found the following:

*48% of pages had some kind of data display,*

*27% of pages had an equation, and*

*7% of pages had both a data display and an equation.*

- Display these results in a Venn diagram.
- What is the probability that a randomly selected sample page had neither a data display nor an equation?
- What is the probability that a randomly selected sample page had a data display but no equation?

## STEP-BY-STEP EXAMPLE

## Using the General Addition Rule

Police report that 78% of drivers stopped on suspicion of drunk driving are given a breath test, 36% a blood test, and 22% both tests.

**Question:** What is the probability that a randomly selected DWI suspect is given

1. a test?
2. a blood test or a breath test, but not both?
3. neither test?



**Plan** Define the events we're interested in. There are no conditions to check; the General Addition Rule works for any events!

**Plot** Make a picture, and use the given probabilities to find the probability for each region.

The blue region represents **A** but not **B**. The green intersection region represents **A** and **B**. Note that since  $P(\mathbf{A}) = 0.78$  and  $P(\mathbf{A} \cap \mathbf{B}) = 0.22$ , the probability of **A** but not **B** must be  $0.78 - 0.22 = 0.56$ .

The yellow region is **B** but not **A**.

The gray region outside both circles represents the outcome neither **A** nor **B**. All the probabilities must total 1, so you can determine the probability of that region by subtraction.

Now, figure out what you want to know. The probabilities can come from the diagram or a formula. Sometimes translating the words to equations is the trickiest step.

Let  $\mathbf{A} = \{\text{suspect is given a breath test}\}$ .

Let  $\mathbf{B} = \{\text{suspect is given a blood test}\}$ .

I know that  $P(\mathbf{A}) = 0.78$

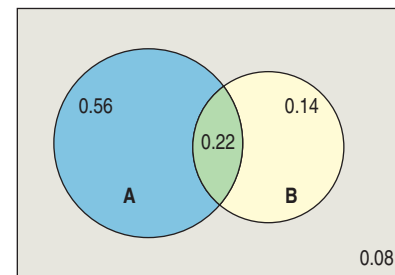
$$P(\mathbf{B}) = 0.36$$

$$P(\mathbf{A} \cap \mathbf{B}) = 0.22$$

$$\text{So } P(\mathbf{A} \cap \mathbf{B}^c) = 0.78 - 0.22 = 0.56$$

$$P(\mathbf{B} \cap \mathbf{A}^c) = 0.36 - 0.22 = 0.14$$

$$\begin{aligned} P(\mathbf{A}^c \cap \mathbf{B}^c) &= 1 - (0.56 + 0.22 + 0.14) \\ &= 0.08 \end{aligned}$$



**Question 1.** What is the probability that the suspect is given a test?



**Mechanics** The probability the suspect is given a test is  $P(\mathbf{A} \cup \mathbf{B})$ . We can use the General Addition Rule, or we can add the probabilities seen in the diagram.

$$\begin{aligned} P(\mathbf{A} \cup \mathbf{B}) &= P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B}) \\ &= 0.78 + 0.36 - 0.22 \\ &= 0.92 \end{aligned}$$

OR

$$P(\mathbf{A} \cup \mathbf{B}) = 0.56 + 0.22 + 0.14 = 0.92$$



**Conclusion** Don't forget to interpret your result in context.

92% of all suspects are given a test.

**Question 2.** What is the probability that the suspect gets either a blood test or a breath test but NOT both?



**Mechanics** We can use the rule, or just add the appropriate probabilities seen in the Venn diagram.

$$P(\mathbf{A \text{ or } B \text{ but NOT both}}) = P(\mathbf{A \cup B}) - P(\mathbf{A \cap B})$$

$$= 0.92 - 0.22 = 0.70$$

OR

$$P(\mathbf{A \text{ or } B \text{ but NOT both}}) = P(\mathbf{A \cap B^c}) + P(\mathbf{B \cap A^c})$$

$$= 0.56 + 0.14 = 0.70$$



**Conclusion** Interpret your result in context.

70% of the suspects get exactly one of the tests.

**Question 3.** What is the probability that the suspect gets neither test?



**Mechanics** Getting neither test is the complement of getting one or the other. Use the Complement Rule or just notice that “neither test” is represented by the region outside both circles.

$$P(\text{neither test}) = 1 - P(\text{either test})$$

$$= 1 - P(\mathbf{A \cup B})$$

$$= 1 - 0.92 = 0.08$$

OR

$$P(\mathbf{A^c \cap B^c}) = 0.08$$



**Conclusion** Interpret your result in context.

Only 8% of the suspects get no test.

## It Depends . . .

Two psychologists surveyed 478 children in grades 4, 5, and 6 in elementary schools in Michigan. They stratified their sample, drawing roughly 1/3 from rural, 1/3 from suburban, and 1/3 from urban schools. Among other questions, they asked the students whether their primary goal was to get good grades, to be popular, or to be good at sports. One question of interest was whether boys and girls at this age had similar goals.

Here’s a *contingency table* giving counts of the students by their goals and sex:

		Goals			Total
		Grades	Popular	Sports	
Sex	Boy	117	50	60	227
	Girl	130	91	30	251
	Total	247	141	90	478

**Table 15.1**

The distribution of goals for boys and girls.



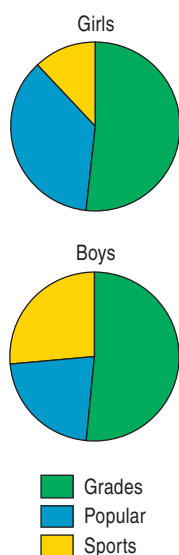


FIGURE 15.1

The distribution of goals for boys and girls.

We looked at contingency tables and graphed *conditional distributions* back in Chapter 3. The pie charts show the *relative frequencies* with which boys and girls named the three goals. It's only a short step from these relative frequencies to probabilities.

Let's focus on this study and make the sample space just the set of these 478 students. If we select a student at random from this study, the probability we select a girl is just the corresponding relative frequency (since we're equally likely to select any of the 478 students). There are 251 girls in the data out of a total of 478, giving a probability of

$$P(\text{girl}) = 251/478 = 0.525$$

The same method works for more complicated events like intersections. For example, what's the probability of selecting a girl whose goal is to be popular? Well, 91 girls named popularity as their goal, so the probability is

$$P(\text{girl} \cap \text{popular}) = 91/478 = 0.190$$

The probability of selecting a student whose goal is to excel at sports is

$$P(\text{sports}) = 90/478 = 0.188$$

What if we are given the information that the selected student is a girl? Would that change the probability that the selected student's goal is sports? You bet it would! The pie charts show that girls are much less likely to say their goal is to excel at sports than are boys. When we restrict our focus to girls, we look only at the girls' row of the table. Of the 251 girls, only 30 of them said their goal was to excel at sports.

We write the probability that a selected student wants to excel at sports *given that we have selected a girl* as

$$P(\text{sports} | \text{girl}) = 30/251 = 0.120$$

For boys, we look at the conditional distribution of goals given "boy" shown in the top row of the table. There, of the 227 boys, 60 said their goal was to excel at sports. So,  $P(\text{sports} | \text{boy}) = 60/227 = 0.264$ , more than twice the girls' probability.

In general, when we want the probability of an event from a *conditional distribution*, we write  $P(\mathbf{B} | \mathbf{A})$  and pronounce it "the probability of  $\mathbf{B}$  given  $\mathbf{A}$ ." A probability that takes into account a given *condition* such as this is called a **conditional probability**.

Let's look at what we did. We worked with the counts, but we could work with the probabilities just as well. There were 30 students who both were girls and had sports as their goal, and there are 251 girls. So we found the probability to be  $30/251$ . To find the probability of the event  $\mathbf{B}$  given the event  $\mathbf{A}$ , we restrict our attention to the outcomes in  $\mathbf{A}$ . We then find in what fraction of *those* outcomes  $\mathbf{B}$  also occurred. Formally, we write:

$$P(\mathbf{B} | \mathbf{A}) = \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{A})}$$

Thinking this through, we can see that it's just what we've been doing, but now with probabilities rather than with counts. Look back at the girls for whom sports was the goal. How did we calculate  $P(\text{sports} | \text{girl})$ ?

The rule says to use probabilities. It says to find  $P(\mathbf{A} \cap \mathbf{B})/P(\mathbf{A})$ . The result is the same whether we use counts or probabilities because the total number in the sample cancels out:

$$\frac{P(\text{sports} \cap \text{girl})}{P(\text{girl})} = \frac{30/478}{251/478} = \frac{30}{251}$$

**AS** **Activity: Birthweights and Smoking.** Does smoking increase the chance of having a baby with low birth weight?

#### NOTATION ALERT:

$P(\mathbf{B} | \mathbf{A})$  is the conditional probability of  $\mathbf{B}$  given  $\mathbf{A}$ .

**AS** **Activity: Conditional Probability.** Simulation is great for seeing conditional probabilities at work.

To use the formula for conditional probability, we're supposed to insist on one restriction. The formula doesn't work if  $P(A)$  is 0. After all, we can't be "given" the fact that **A** was true if the probability of **A** is 0!

Let's take our rule out for a spin. What's the probability that we have selected a girl *given* that the selected student's goal is popularity? Applying the rule, we get

$$\begin{aligned} P(\text{girl} \mid \text{popular}) &= \frac{P(\text{girl} \cap \text{popular})}{P(\text{popular})} \\ &= \frac{91/478}{141/478} = \frac{91}{141}. \end{aligned}$$

### FOR EXAMPLE

#### Finding a conditional probability

**Recap:** Our survey found that 56% of college students live on campus, 62% have a campus meal program, and 42% do both.

**Question:** While dining in a campus facility open only to students with meal plans, you meet someone interesting. What is the probability that your new acquaintance lives on campus?

Let  $L = \{\text{student lives on campus}\}$  and  $M = \{\text{student has a campus meal plan}\}$ .

$$\begin{aligned} P(\text{student lives on campus given that the student has a meal plan}) &= P(L \mid M) \\ &= \frac{P(L \cap M)}{P(M)} \\ &= \frac{0.42}{0.62} \\ &\approx 0.677 \end{aligned}$$

There's a probability of about 0.677 that a student with a meal plan lives on campus.

## The General Multiplication Rule

Remember the Multiplication Rule for the probability of **A** and **B**? It said

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}) \text{ when } \mathbf{A} \text{ and } \mathbf{B} \text{ are independent.}$$

Now we can write a more general rule that doesn't require independence. In fact, we've *already* written it down. We just need to rearrange the equation a bit.

The equation in the definition for conditional probability contains the probability of **A** and **B**. Rewriting the equation gives

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} \mid \mathbf{A}).$$

This is a **General Multiplication Rule** for compound events that does not require the events to be independent. Better than that, it even makes sense. The probability that two events, **A** and **B**, *both* occur is the probability that event **A** occurs multiplied by the probability that event **B** *also* occurs—that is, by the probability that event **B** occurs *given* that event **A** occurs.

Of course, there's nothing special about which set we call **A** and which one we call **B**. We should be able to state this the other way around. And indeed we can. It is equally true that

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{B}) \times P(\mathbf{A} \mid \mathbf{B}).$$

**AS** **Activity: The General Multiplication Rule.** The best way to understand the General Multiplication Rule is with an experiment.

# Independence

If we had to pick one idea in this chapter that you should understand and remember, it's the definition and meaning of independence. We'll need this idea in every one of the chapters that follow.

**AS**

**Activity: Independence.**

Are *Smoking and Low Birthweight* independent?

In earlier chapters we said informally that two events were independent if learning that one occurred didn't change what you thought about the other occurring. Now we can be more formal. Events **A** and **B** are independent if (and only if) the probability of **A** is the same when we are given that **B** has occurred. That is,  $P(\mathbf{A}) = P(\mathbf{A} | \mathbf{B})$ .

Although sometimes your intuition is enough, now that we have the formal rule, use it whenever you can.

Let's return to the question of just what it means for events to be independent. We've said informally that what we mean by independence is that the outcome of one event does not influence the probability of the other. With our new notation for conditional probabilities, we can write a formal definition: **Events **A** and **B** are independent whenever**

$$P(\mathbf{B} | \mathbf{A}) = P(\mathbf{B}).$$

Now we can see that the Multiplication Rule for independent events we saw in Chapter 14 is just a special case of the General Multiplication Rule. The general rule says

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} | \mathbf{A}).$$

whether the events are independent or not. But when events **A** and **B** are independent, we can write  $P(\mathbf{B})$  for  $P(\mathbf{B} | \mathbf{A})$  and we get back our simple rule:

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}).$$

Sometimes people use this statement as the definition of independent events, but we find the other definition more intuitive. Either way, the idea is that for independent events, the probability of one doesn't change when the other occurs.

Is the probability of having good grades as a goal independent of the sex of the responding student? Looks like it might be. We need to check whether

$$\begin{aligned} P(\text{grades} | \text{girl}) &= P(\text{grades}) \\ \frac{130}{251} &= 0.52 \stackrel{?}{=} \frac{247}{478} = 0.52 \end{aligned}$$

To two decimal place accuracy, it looks like we can consider choosing good grades as a goal to be independent of sex.

On the other hand,  $P(\text{sports})$  is  $90/478$ , or about 18.8%, but  $P(\text{sports} | \text{boy})$  is  $60/227 = 26.4\%$ . Because these probabilities aren't equal, we can be pretty sure that choosing success in sports as a goal is not independent of the student's sex.

## FOR EXAMPLE

### Checking for independence

**Recap:** Our survey told us that 56% of college students live on campus, 62% have a campus meal program, and 42% do both.

**Question:** Are living on campus and having a meal plan independent? Are they disjoint?

Let  $\mathbf{L} = \{\text{student lives on campus}\}$  and  $\mathbf{M} = \{\text{student has a campus meal plan}\}$ . If these events are independent, then knowing that a student lives on campus doesn't affect the probability that he or she has a meal plan. I'll check to see if  $P(\mathbf{M} | \mathbf{L}) = P(\mathbf{M})$ :

$$\begin{aligned} P(\mathbf{M} | \mathbf{L}) &= \frac{P(\mathbf{L} \cap \mathbf{M})}{P(\mathbf{L})} \\ &= \frac{0.42}{0.56} \\ &= 0.75, \quad \text{but } P(\mathbf{M}) = 0.62. \end{aligned}$$

Because  $0.75 \neq 0.62$ , the events are not independent; students who live on campus are more likely to have meal plans. Living on campus and having a meal plan are not disjoint either; in fact, 42% of college students do both.

## Independent $\neq$ Disjoint

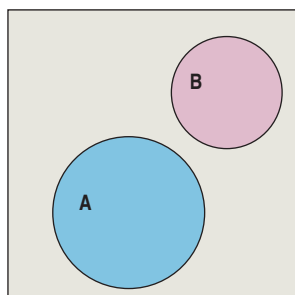


FIGURE 15.2

Because these events are mutually exclusive, learning that **A** happened tells us that **B** didn't. The probability of **B** has changed from whatever it was to zero. So the disjoint events **A** and **B** are not independent.

Are disjoint events independent? These concepts seem to have similar ideas of separation and distinctness about them, but in fact disjoint events *cannot* be independent.<sup>2</sup> Let's see why. Consider the two disjoint events {you get an A in this course} and {you get a B in this course}. They're disjoint because they have no outcomes in common. Suppose you learn that you *did* get an A in the course. Now what is the probability that you got a B? You can't get both grades, so it must be 0.

Think about what that means. Knowing that the first event (getting an A) occurred changed your probability for the second event (down to 0). So these events aren't independent.

Mutually exclusive events can't be independent. They have no outcomes in common, so if one occurs, the other doesn't. A common error is to treat disjoint events as if they were independent and apply the Multiplication Rule for independent events. Don't make that mistake.



### JUST CHECKING

2. The American Association for Public Opinion Research (AAPOR) is an association of about 1600 individuals who share an interest in public opinion and survey research. They report that typically as few as 10% of random phone calls result in a completed interview. Reasons are varied, but some of the most common include no answer, refusal to cooperate, and failure to complete the call.

Which of the following events are independent, which are disjoint, and which are neither independent nor disjoint?

- a) **A** = Your telephone number is randomly selected. **B** = You're not at home at dinnertime when they call.
- b) **A** = As a selected subject, you complete the interview. **B** = As a selected subject, you refuse to cooperate.
- c) **A** = You are not at home when they call at 11 a.m. **B** = You are employed full-time.

## Depending on Independence

**A S** **Video: Is There a Hot Hand in Basketball?** Most coaches and fans believe that basketball players sometimes get "hot" and make more of their shots. What do the conditional probabilities say?

**A S** **Activity: Hot Hand Simulation.** Can you tell the difference between real and simulated sequences of basketball shot hits and misses?

It's much easier to think about independent events than to deal with conditional probabilities. It seems that most people's natural intuition for probabilities breaks down when it comes to conditional probabilities. Someone may estimate the probability of a compound event by multiplying the probabilities of its component events together without asking seriously whether those probabilities are independent.

For example, experts have assured us that the probability of a major commercial nuclear plant failure is so small that we should not expect such a failure to occur even in a span of hundreds of years. After only a few decades of commercial nuclear power, however, the world has seen two failures (Chernobyl and Three Mile Island). How could the estimates have been so wrong?

<sup>2</sup> Well, technically two disjoint events *can* be independent, but only if the probability of one of the events is 0. For practical purposes, though, we can ignore this case. After all, as statisticians we don't anticipate having data about things that never happen.

One simple part of the failure calculation is to test a particular valve and determine that valves such as this one fail only once in, say, 100 years of normal use. For a coolant failure to occur, several valves must fail. So we need the compound probability,  $P(\text{valve 1 fails and valve 2 fails and } \dots)$ . A simple risk assessment might multiply the small probability of one valve failure together as many times as needed.

But if the valves all came from the same manufacturer, a flaw in one might be found in the others. And maybe when the first fails, it puts additional pressure on the next one in line. In either case, the events aren't independent and so we can't simply multiply the probabilities together.

Whenever you see probabilities multiplied together, stop and ask whether you think they are really independent.

## Tables and Conditional Probability

One of the easiest ways to think about conditional probabilities is with contingency tables. We did that earlier in the chapter when we began our discussion. But sometimes we're given probabilities without a table. You can often construct a simple table to correspond to the probabilities.

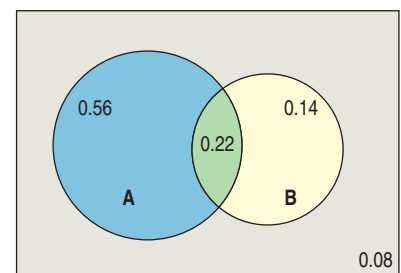
For instance, in the drunk driving example, we were told that 78% of suspect drivers get a breath test, 36% a blood test, and 22% both. That's enough information. Translating percentages to probabilities, what we know looks like this:

		Breath Test		
		Yes	No	Total
Blood Test	Yes	0.22		0.36
	No			
	Total	0.78		1.00

Notice that the 0.78 and 0.36 are *marginal* probabilities and so they go into the *margins*. The 0.22 is the probability of getting both tests—a breath test *and* a blood test—so that's a *joint* probability. Those belong in the interior of the table.

Because the cells of the table show disjoint events, the probabilities always add to the marginal totals going across rows or down columns. So, filling in the rest of the table is quick:

		Breath Test		
		Yes	No	Total
Blood Test	Yes	0.22	0.14	0.36
	No	0.56	0.08	0.64
	Total	0.78	0.22	1.00



Compare this with the Venn diagram. Notice which entries in the table match up with the sets in this diagram. Whether a Venn diagram or a table is better to use will depend on what you are given and the questions you're being asked. Try both.

## STEP-BY-STEP EXAMPLE

## Are the Events Disjoint? Independent?

Let's take another look at the drunk driving situation. Police report that 78% of drivers are given a breath test, 36% a blood test, and 22% both tests.

**Questions:** 1. Are giving a DWI suspect a blood test and a breath test mutually exclusive?  
2. Are giving the two tests independent?

THINK

**Plan** Define the events we're interested in.  
State the given probabilities.

Let  $A = \{\text{suspect is given a breath test}\}$

Let  $B = \{\text{suspect is given a blood test}\}$ .

I know that  $P(A) = 0.78$

$P(B) = 0.36$

$P(A \cap B) = 0.22$

**Question 1.** Are giving a DWI suspect a blood test and a breath test mutually exclusive?

SHOW

**Mechanics** Disjoint events cannot *both* happen at the same time, so check to see if  $P(A \cap B) = 0$ .

$P(A \cap B) = 0.22$ . Since some suspects are given both tests,  $P(A \cap B) \neq 0$ . The events are not mutually exclusive.

TELL

**Conclusion** State your conclusion in context.

22% of all suspects get both tests, so a breath test and a blood test are not disjoint events.

**Question 2.** Are the two tests independent?

THINK

**Plan** Make a table.

		Breath Test		Total
		Yes	No	
Blood Test	Yes	0.22	0.14	0.36
	No	0.56	0.08	0.64
	Total	0.78	0.22	1.00

SHOW

**Mechanics** Does getting a breath test change the probability of getting a blood test? That is, does  $P(B|A) = P(B)$ ?

Because the two probabilities are *not* the same, the events are not independent.

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.22}{0.78} \approx 0.28$$

$$P(B) = 0.36$$

$$P(B|A) \neq P(B)$$



**Conclusion** Interpret your results in context.

Overall, 36% of the drivers get blood tests, but only 28% of those who get a breath test do. Since suspects who get a breath test are less likely to have a blood test, the two events are not independent.



### JUST CHECKING

3. Remember our sample of pages in this book from the earlier Just Checking . . . ?

48% of pages had a data display.

27% of pages had an equation, and

7% of pages had both a data display and an equation.

- Make a contingency table for the variables *display* and *equation*.
- What is the probability that a randomly selected sample page with an equation also had a data display?
- Are having an equation and having a data display disjoint events?
- Are having an equation and having a data display independent events?

## Drawing Without Replacement

Room draw is a process for assigning dormitory rooms to students who live on campus. Sometimes, when students have equal priority, they are randomly assigned to the currently available dorm rooms. When it's time for you and your friend to draw, there are 12 rooms left. Three are in Gold Hall, a very desirable dorm with spacious wood-paneled rooms. Four are in Silver Hall, centrally located but not quite as desirable. And five are in Wood Hall, a new dorm with cramped rooms, located half a mile from the center of campus on the edge of the woods.

You get to draw first, and then your friend will draw. Naturally, you would both like to score rooms in Gold. What are your chances? In particular, what's the chance that you *both* can get rooms in Gold?

When you go first, the chance that *you* will draw one of the Gold rooms is  $3/12$ . Suppose you do. Now, with you clutching your prized room assignment, what chance does your friend have? At this point there are only 11 rooms left and just 2 left in Gold, so your friend's chance is now  $2/11$ .

Using our notation, we write

$$P(\text{friend draws Gold} \mid \text{you draw Gold}) = 2/11.$$

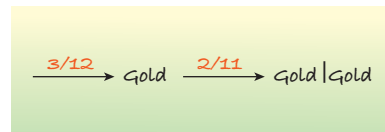
The reason the denominator changes is that we draw these rooms *without replacement*. That is, once one is drawn, it doesn't go back into the pool.

We often sample without replacement. When we draw from a very large population, the change in the denominator is too small to worry about. But when there's a small population to draw from, as in this case, we need to take note and adjust the probabilities.

What are the chances that *both* of you will luck out? Well, now we've calculated the two probabilities we need for the General Multiplication Rule, so we can write:

$$\begin{aligned} &P(\text{you draw Gold} \cap \text{friend draws Gold}) \\ &= P(\text{you draw Gold}) \times P(\text{friend draws Gold} \mid \text{you draw Gold}) \\ &= 3/12 \times 2/11 = 1/22 = 0.045 \end{aligned}$$

In this instance, it doesn't matter who went first, or even if the rooms were drawn simultaneously. Even if the room draw was accomplished by shuffling cards containing the names of the dormitories and then dealing them out to 12 applicants (rather than by each student drawing a room in turn), we can still *think* of the calculation as having taken place in two steps:



Diagramming conditional probabilities leads to a more general way of helping us think with pictures—one that works for calculating conditional probabilities even when they involve different variables.

## Tree Diagrams

For men, binge drinking is defined as having five or more drinks in a row, and for women as having four or more drinks in a row. (The difference is because of the average difference in weight.) According to a study by the Harvard School of Public Health (H. Wechsler, G. W. Dowdall, A. Davenport, and W. DeJong, "Binge Drinking on Campus: Results of a National Study"), 44% of college students engage in binge drinking, 37% drink moderately, and 19% abstain entirely. Another study, published in the *American Journal of Health Behavior*, finds that among binge drinkers aged 21 to 34, 17% have been involved in an alcohol-related automobile accident, while among non-bingers of the same age, only 9% have been involved in such accidents.

What's the probability that a randomly selected college student will be a binge drinker who has had an alcohol-related car accident?

To start, we see that the probability of selecting a binge drinker is about 44%. To find the probability of selecting someone who is both a binge drinker and a driver with an alcohol-related accident, we would need to pull out the General Multiplication Rule and multiply the probability of one of the events by the conditional probability of the other given the first.

Or we *could* make a picture. Which would you prefer?

We thought so.

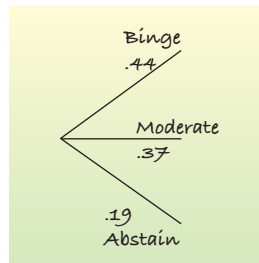
The kind of picture that helps us think through this kind of reasoning is called a **tree diagram**, because it shows sequences of events, like those we had in room draw, as paths that look like branches of a tree. It is a good idea to make a tree diagram almost any time you plan to use the General Multiplication Rule. The number of different paths we can take can get large, so we usually draw the tree starting from the left and growing vine-like across the page, although sometimes you'll see them drawn from the bottom up or top down.

"Why," said the Dodo, "the best way to explain it is to do it."

—Lewis Carroll



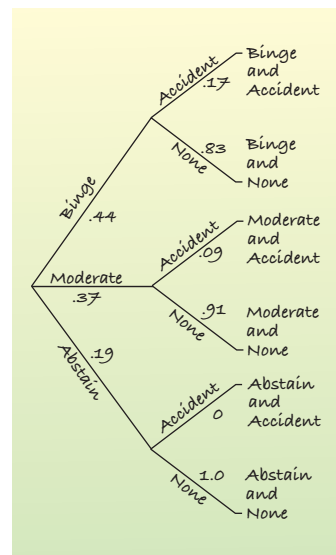
The first branch of our tree separates students according to their drinking habits. We label each branch of the tree with a possible outcome and its corresponding probability.



**FIGURE 15.3**

We can diagram the three outcomes of drinking and indicate their respective probabilities with a simple tree diagram.

Notice that we cover all possible outcomes with the branches. The probabilities add up to one. But we're also interested in car accidents. The probability of having an alcohol-related accident *depends* on one's drinking behavior. Because the probabilities are *conditional*, we draw the alternatives separately on each branch of the tree:



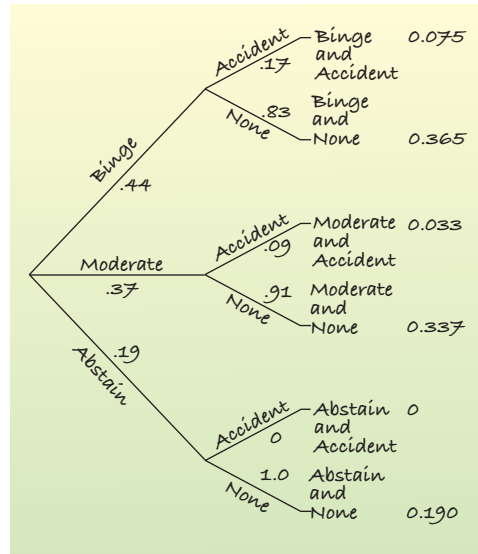
**FIGURE 15.4**

Extending the tree diagram, we can show both drinking and accident outcomes. The accident probabilities are conditional on the drinking outcomes, and they change depending on which branch we follow. Because we are concerned only with alcohol-related accidents, the conditional probability  $P(\text{accident} | \text{abstinence})$  must be 0.

On each of the second set of branches, we write the possible outcomes associated with having an alcohol-related car accident (having an accident or not) and the associated probability. These probabilities are different because they are *conditional* depending on the student's drinking behavior. (It shouldn't be too surprising that those who binge drink have a higher probability of alcohol-related accidents.) The probabilities add up to one, because given the outcome on the first branch, these outcomes cover all the possibilities. Looking back at the General Multiplication Rule, we can see how the tree depicts the calculation. To find the probability that a randomly selected student will be a binge drinker who has had an alcohol-related car accident, we follow the top branches. The probability of selecting a binger is 0.44. The conditional probability of an accident *given* binge drinking is 0.17. The General Multiplication Rule tells us that to find the *joint* probability of being a binge drinker and having an accident, we multiply these two probabilities together:

$$\begin{aligned} P(\text{binge} \cap \text{accident}) &= P(\text{binge}) \times P(\text{accident} | \text{binge}) \\ &= 0.44 \times 0.17 = 0.075 \end{aligned}$$

And we can do the same for each combination of outcomes:



**FIGURE 15.5**

We can find the probabilities of compound events by multiplying the probabilities along the branch of the tree that leads to the event, just the way the General Multiplication Rule specifies.

The probability of abstaining and having an alcohol-related accident is, of course, zero.

All the outcomes at the far right are disjoint because at each branch of the tree we chose between disjoint alternatives. And they are *all* the possibilities, so the probabilities on the far right must add up to one. Always check!

Because the final outcomes are disjoint, we can add up their probabilities to get probabilities for compound events. For example, what’s the probability that a selected student has had an alcohol-related car accident? We simply find *all* the outcomes on the far right in which an accident has happened. There are three and we can add their probabilities:  $0.075 + 0.033 + 0 = 0.108$ —almost an 11% chance.

## Reversing the Conditioning

If we know a student has had an alcohol-related accident, what’s the probability that the student is a binge drinker? That’s an interesting question, but we can’t just read it from the tree. The tree gives us  $P(\text{accident} | \text{binge})$ , but we want  $P(\text{binge} | \text{accident})$ —conditioning in the other direction. The two probabilities are definitely *not* the same. We have reversed the conditioning.

We may not have the conditional probability we want, but we do know everything we need to know to find it. To find a conditional probability, we need the probability that both events happen divided by the probability that the given event occurs. We have already found the probability of an alcohol-related accident:  $0.075 + 0.033 + 0 = 0.108$ .

The joint probability that a student is both a binge drinker and someone who’s had an alcohol-related accident is found at the top branch: 0.075. We’ve restricted the *Who* of the problem to the students with alcohol-related accidents, so we divide the two to find the conditional probability:

$$\begin{aligned}
 P(\text{binge} | \text{accident}) &= \frac{P(\text{binge} \cap \text{accident})}{P(\text{accident})} \\
 &= \frac{0.075}{0.108} = 0.694
 \end{aligned}$$

The chance that a student who has an alcohol-related car accident is a binge drinker is more than 69%! As we said, reversing the conditioning is rarely intuitive, but tree diagrams help us keep track of the calculation when there aren’t too many alternatives to consider.

## STEP-BY-STEP EXAMPLE

## Reversing the Conditioning

When the authors were in college, there were only three requirements for graduation that were the same for all students: You had to be able to tread water for 2 minutes, you had to learn a foreign language, and you had to be free of tuberculosis. For the last requirement, all freshmen had to take a TB screening test that consisted of a nurse jabbing what looked like a corncob holder into your forearm. You were then expected to report back in 48 hours to have it checked. If you were healthy and TB-free, your arm was supposed to look as though you'd never had the test.

Sometime during the 48 hours, one of us had a reaction. When he finally saw the nurse, his arm was about 50% bigger than normal and a very unhealthy red. Did he have TB? The nurse had said that the test was about 99% effective, so it seemed that the chances must be pretty high that he had TB. How high do you think the chances were? Go ahead and guess. Guess low.

We'll call **TB** the event of actually having TB and **+** the event of testing positive. To start a tree, we need to know  $P(\text{TB})$ , the probability of having TB.<sup>3</sup> We also need to know the conditional probabilities  $P(+|\text{TB})$  and  $P(+|\text{TB}^c)$ . Diagnostic tests can make two kinds of errors. They can give a positive result for a healthy person (a *false positive*) or a negative result for a sick person (a *false negative*). Being 99% accurate usually means a false-positive rate of 1%. That is, someone who doesn't have the disease has a 1% chance of testing positive anyway. We can write  $P(+|\text{TB}^c) = 0.01$ .

Since a false negative is more serious (because a sick person might not get treatment), tests are usually constructed to have a lower false-negative rate. We don't know exactly, but let's assume a 0.1% false-negative rate. So only 0.1% of sick people test negative. We can write  $P(-|\text{TB}) = 0.001$ .

## THINK

**Plan** Define the events we're interested in and their probabilities.

Figure out what you want to know in terms of the events. Use the notation of conditional probability to write the event whose probability you want to find.

Let  $\text{TB} = \{\text{having TB}\}$  and  $\text{TB}^c = \{\text{no TB}\}$   
 $+$  = {testing positive} and  
 $-$  = {testing negative}

I know that  $P(+|\text{TB}^c) = 0.01$  and  
 $P(-|\text{TB}) = 0.001$ . I also know that  
 $P(\text{TB}) = 0.00005$ .

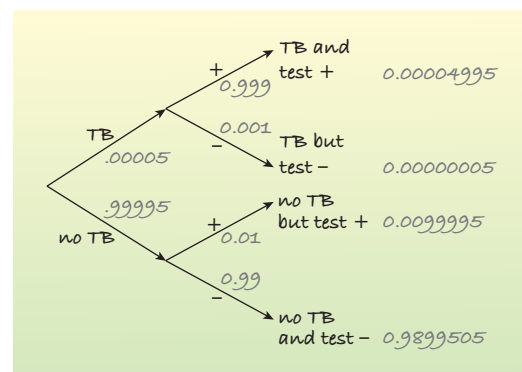
I'm interested in the probability that the author had TB given that he tested positive:  $P(\text{TB} | +)$ .

## SHOW

**Plot** Draw the tree diagram. When probabilities are very small like these are, be careful to keep all the significant digits.

To finish the tree we need  $P(\text{TB}^c)$ ,  $P(-|\text{TB}^c)$ , and  $P(-|\text{TB})$ . We can find each of these from the Complement Rule:

$$\begin{aligned} P(\text{TB}^c) &= 1 - P(\text{TB}) = 0.99995 \\ P(-|\text{TB}^c) &= 1 - P(+|\text{TB}^c) \\ &= 1 - 0.01 = 0.99 \text{ and} \\ P(+|\text{TB}) &= 1 - P(-|\text{TB}) \\ &= 1 - 0.001 = 0.999 \end{aligned}$$



<sup>3</sup> This isn't given, so we looked it up. Although TB is a matter of serious concern to public health officials, it is a fairly uncommon disease, with an incidence of about 5 cases per 100,000 in the United States (see <http://www.cdc.gov/tb/default.htm>).

**Mechanics** Multiply along the branches to find the probabilities of the four possible outcomes. Check your work by seeing if they total 1.

Add up the probabilities corresponding to the condition of interest—in this case, testing positive. We can add because the tree shows disjoint events.

Divide the probability of both events occurring (here, having TB and a positive test) by the probability of satisfying the condition (testing positive).

(Check:  $0.00004995 + 0.00000005 + 0.00999995 + 0.989995050 = 1$ )

$$\begin{aligned} P(+) &= P(\text{TB} \cap +) + P(\text{TB}^c \cap +) \\ P &= 0.00004995 + 0.00999995 \\ &= 0.01004945 \end{aligned}$$

$$\begin{aligned} P(\text{TB} | +) &= \frac{P(\text{TB} \cap +)}{P(+)} \\ &= \frac{0.00004995}{0.01004945} \\ &= 0.00497 \end{aligned}$$



**Conclusion** Interpret your result in context.

The chance of having TB after you test positive is less than 0.5%.

When we reverse the order of conditioning, we change the *Who* we are concerned with. With events of low probability, the result can be surprising. That's the reason patients who test positive for HIV, for example, are always told to seek medical counseling. They may have only a small chance of actually being infected. That's why global drug or disease testing can have unexpected consequences if people interpret *testing positive* as *being positive*.

## Bayes's Rule



The Reverend Thomas Bayes is credited posthumously with the rule that is the foundation of Bayesian Statistics.

When we have  $P(\mathbf{A} | \mathbf{B})$  but want the *reverse* probability  $P(\mathbf{B} | \mathbf{A})$ , we need to find  $P(\mathbf{A} \cap \mathbf{B})$  and  $P(\mathbf{A})$ . A tree is often a convenient way of finding these probabilities. It can work even when we have more than two possible events, as we saw in the binge-drinking example. Instead of using the tree, we *could* write the calculation algebraically, showing exactly how we found the quantities that we needed:  $P(\mathbf{A} \cap \mathbf{B})$  and  $P(\mathbf{A})$ . The result is a formula known as Bayes's Rule, after the Reverend Thomas Bayes (1702?–1761), who was credited with the rule after his death, when he could no longer defend himself. Bayes's Rule is quite important in Statistics and is the foundation of an approach to Statistical analysis known as Bayesian Statistics. Although the simple rule deals with two alternative outcomes, the rule can be extended to the situation in which there are more than two branches to the first split of the tree. The principle remains the same (although the math gets more difficult). Bayes's Rule is just a formula<sup>4</sup> for reversing the probability from the conditional probability that you're originally given, the same feat we accomplished with our tree diagram.

<sup>4</sup> Bayes's Rule for two events says that  $P(\mathbf{B} | \mathbf{A}) = \frac{P(\mathbf{A} | \mathbf{B})P(\mathbf{B})}{P(\mathbf{A} | \mathbf{B})P(\mathbf{B}) + P(\mathbf{A} | \mathbf{B}^c)P(\mathbf{B}^c)}$ .

Masochists may wish to try it with the TB testing probabilities. (It's easier to just draw the tree, isn't it?)

## FOR EXAMPLE

## Reversing the conditioning

A recent Maryland highway safety study found that in 77% of all accidents the driver was wearing a seatbelt. Accident reports indicated that 92% of those drivers escaped serious injury (defined as hospitalization or death), but only 63% of the non-belted drivers were so fortunate.

**Question:** What's the probability that a driver who was seriously injured wasn't wearing a seatbelt?

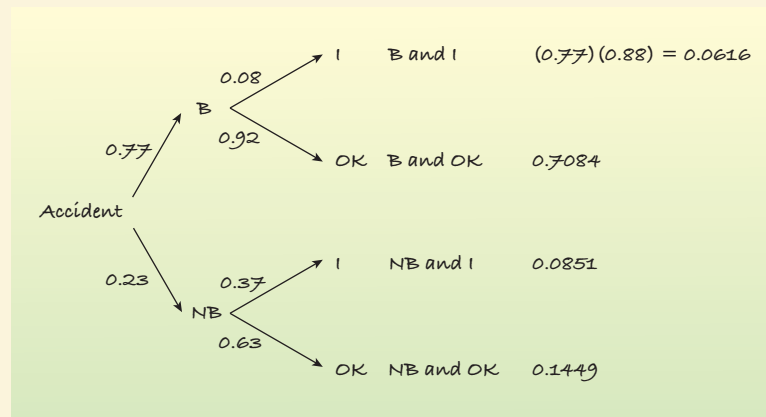
Let  $B$  = the driver was wearing a seatbelt, and  $NB$  = no belt.

Let  $I$  = serious injury or death, and  $OK$  = not seriously injured.

I know  $P(B) = 0.77$ , so  $P(NB) = 1 - 0.77 = 0.23$ .

Also,  $P(OK|B) = 0.92$ , so  $P(I|B) = 0.08$

and  $P(OK|NB) = 0.63$ , so  $P(I|NB) = 0.37$



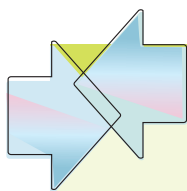
$$P(NB|I) = \frac{P(NB \text{ and } I)}{P(I)} = \frac{0.0851}{0.0616 + 0.0851} = 0.58$$

Even though only 23% of drivers weren't wearing seatbelts, they accounted for 58% of all the deaths and serious injuries.

Just some advice from your friends, the authors: *Please buckle up!* (We want you to finish this course.)

## WHAT CAN GO WRONG?

- ▶ **Don't use a simple probability rule where a general rule is appropriate.** Don't assume independence without reason to believe it. Don't assume that outcomes are disjoint without checking that they are. Remember that the general rules always apply, even when outcomes are in fact independent or disjoint.
- ▶ **Don't find probabilities for samples drawn without replacement as if they had been drawn with replacement.** Remember to adjust the denominator of your probabilities. This warning applies only when we draw from small populations or draw a large fraction of a finite population. When the population is very large relative to the sample size, the adjustments make very little difference, and we ignore them.
- ▶ **Don't reverse conditioning naively.** As we have seen, the probability of  $A$  given  $B$  may not, and, in general does not, resemble the probability of  $B$  given  $A$ . The true probability may be counterintuitive.
- ▶ **Don't confuse "disjoint" with "independent."** Disjoint events *cannot* happen at the same time. When one happens, you know the other did not, so  $P(B|A) = 0$ . Independent events *must* be able to happen at the same time. When one happens, you know it has no effect on the other, so  $P(B|A) = P(B)$ .

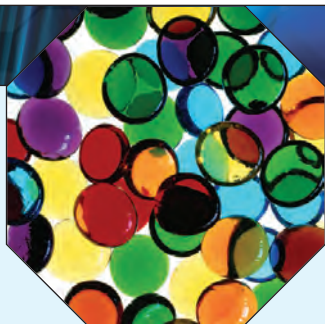


## CONNECTIONS

This chapter shows the unintuitive side of probability. If you've been thinking, "My mind doesn't work this way," you're probably right. Humans don't seem to find conditional and compound probabilities natural and often have trouble with them. Even statisticians make mistakes with conditional probability.

Our central connection is to the guiding principle that Statistics is about understanding the world. The events discussed in this chapter are close to the kinds of real-world situations in which understanding probabilities matters. The methods and concepts of this chapter are the tools you need to understand the part of the real world that deals with the outcomes of complex, uncertain events.

## WHAT HAVE WE LEARNED?



The last chapter's basic rules of probability are important, but they work only in special cases—when events are disjoint or independent. Now we've learned the more versatile General Addition Rule and General Multiplication Rule. We've also learned about conditional probabilities, and seen that reversing the conditioning can give surprising results.

We've learned the value of Venn diagrams, tables, and tree diagrams to help organize our thinking about probabilities.

Most important, we've learned to think clearly about independence. We've seen how to use conditional probability to determine whether two events are independent and to work with events that are not independent. A sound understanding of independence will be important throughout the rest of this book.

### Terms

General Addition Rule

343. For any two events, **A** and **B**, the probability of **A** or **B** is

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B}).$$

Conditional probability

$$347. P(\mathbf{B} | \mathbf{A}) = \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{A})}$$

$P(\mathbf{B} | \mathbf{A})$  is read "the probability of **B** given **A**."

General Multiplication Rule

348. For any two events, **A** and **B**, the probability of **A** and **B** is

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B} | \mathbf{A}).$$

Independence (used formally)

349. Events **A** and **B** are independent when  $P(\mathbf{B} | \mathbf{A}) = P(\mathbf{B})$ .

Tree diagram

354. A display of conditional events or probabilities that is helpful in thinking through conditioning.

### Skills

THINK

► Understand the concept of conditional probability as redefining the *Who* of concern, according to the information about the event that is *given*.

► Understand the concept of independence.

SHOW

► Know how and when to apply the General Addition Rule.

► Know how to find probabilities for compound events as fractions of counts of occurrences in a two-way table.



- ▶ Know how and when to apply the General Multiplication Rule.
- ▶ Know how to make and use a tree diagram to understand conditional probabilities and reverse conditioning.
- ▶ Be able to make a clear statement about a conditional probability that makes clear how the condition affects the probability.
- ▶ Avoid making statements that assume independence of events when there is no clear evidence that they are in fact independent.

## EXERCISES

1. **Homes.** Real estate ads suggest that 64% of homes for sale have garages, 21% have swimming pools, and 17% have both features. What is the probability that a home for sale has
  - a) a pool or a garage?
  - b) neither a pool nor a garage?
  - c) a pool but no garage?
2. **Travel.** Suppose the probability that a U.S. resident has traveled to Canada is 0.18, to Mexico is 0.09, and to both countries is 0.04. What's the probability that an American chosen at random has
  - a) traveled to Canada but not Mexico?
  - b) traveled to either Canada or Mexico?
  - c) not traveled to either country?
3. **Amenities.** A check of dorm rooms on a large college campus revealed that 38% had refrigerators, 52% had TVs, and 21% had both a TV and a refrigerator. What's the probability that a randomly selected dorm room has
  - a) a TV but no refrigerator?
  - b) a TV or a refrigerator, but not both?
  - c) neither a TV nor a refrigerator?
4. **Workers.** Employment data at a large company reveal that 72% of the workers are married, that 44% are college graduates, and that half of the college grads are married. What's the probability that a randomly chosen worker
  - a) is neither married nor a college graduate?
  - b) is married but not a college graduate?
  - c) is married or a college graduate?
5. **Global survey.** The marketing research organization GfK Custom Research North America conducts a yearly survey on consumer attitudes worldwide. They collect demographic information on the roughly 1500 respondents from each country that they survey. Here is a table showing the number of people with various levels of education in five countries:

Educational Level by Country

	Post-graduate	College	Some high school	Primary or less	No answer	Total
China	7	315	671	506	3	1502
France	69	388	766	309	7	1539
India	161	514	622	227	11	1535
U.K.	58	207	1240	32	20	1557
USA	84	486	896	87	4	1557
Total	379	1910	4195	1161	45	7690

If we select someone at random from this survey,

- a) what is the probability that the person is from the United States?
  - b) what is the probability that the person completed his or her education before college?
  - c) what is the probability that the person is from France or did some post-graduate study?
  - d) what is the probability that the person is from France and finished only primary school or less?
6. **Birth order.** A survey of students in a large Introductory Statistics class asked about their birth order (1 = oldest or only child) and which college of the university they were enrolled in. Here are the results:

Birth Order

		1 or only	2 or more	Total
College	Arts & Sciences	34	23	57
	Agriculture	52	41	93
	Human Ecology	15	28	43
	Other	12	18	30
	Total	113	110	223



- ▶ Know how and when to apply the General Multiplication Rule.
- ▶ Know how to make and use a tree diagram to understand conditional probabilities and reverse conditioning.
- ▶ Be able to make a clear statement about a conditional probability that makes clear how the condition affects the probability.
- ▶ Avoid making statements that assume independence of events when there is no clear evidence that they are in fact independent.

## EXERCISES

1. **Homes.** Real estate ads suggest that 64% of homes for sale have garages, 21% have swimming pools, and 17% have both features. What is the probability that a home for sale has
  - a) a pool or a garage?
  - b) neither a pool nor a garage?
  - c) a pool but no garage?
2. **Travel.** Suppose the probability that a U.S. resident has traveled to Canada is 0.18, to Mexico is 0.09, and to both countries is 0.04. What's the probability that an American chosen at random has
  - a) traveled to Canada but not Mexico?
  - b) traveled to either Canada or Mexico?
  - c) not traveled to either country?
3. **Amenities.** A check of dorm rooms on a large college campus revealed that 38% had refrigerators, 52% had TVs, and 21% had both a TV and a refrigerator. What's the probability that a randomly selected dorm room has
  - a) a TV but no refrigerator?
  - b) a TV or a refrigerator, but not both?
  - c) neither a TV nor a refrigerator?
4. **Workers.** Employment data at a large company reveal that 72% of the workers are married, that 44% are college graduates, and that half of the college grads are married. What's the probability that a randomly chosen worker
  - a) is neither married nor a college graduate?
  - b) is married but not a college graduate?
  - c) is married or a college graduate?
5. **Global survey.** The marketing research organization GfK Custom Research North America conducts a yearly survey on consumer attitudes worldwide. They collect demographic information on the roughly 1500 respondents from each country that they survey. Here is a table showing the number of people with various levels of education in five countries:

Educational Level by Country

	Post-graduate	College	Some high school	Primary or less	No answer	Total
China	7	315	671	506	3	1502
France	69	388	766	309	7	1539
India	161	514	622	227	11	1535
U.K.	58	207	1240	32	20	1557
USA	84	486	896	87	4	1557
Total	379	1910	4195	1161	45	7690

If we select someone at random from this survey,

- a) what is the probability that the person is from the United States?
  - b) what is the probability that the person completed his or her education before college?
  - c) what is the probability that the person is from France or did some post-graduate study?
  - d) what is the probability that the person is from France and finished only primary school or less?
6. **Birth order.** A survey of students in a large Introductory Statistics class asked about their birth order (1 = oldest or only child) and which college of the university they were enrolled in. Here are the results:

Birth Order

		1 or only	2 or more	Total
College	Arts & Sciences	34	23	57
	Agriculture	52	41	93
	Human Ecology	15	28	43
	Other	12	18	30
	Total	113	110	223



Suppose we select a student at random from this class.

What is the probability that the person is

- a Human Ecology student?
  - a firstborn student?
  - firstborn *and* a Human Ecology student?
  - firstborn *or* a Human Ecology student?
7. **Cards.** You draw a card at random from a standard deck of 52 cards. Find each of the following conditional probabilities:
- The card is a heart, given that it is red.
  - The card is red, given that it is a heart.
  - The card is an ace, given that it is red.
  - The card is a queen, given that it is a face card.
8. **Pets.** In its monthly report, the local animal shelter states that it currently has 24 dogs and 18 cats available for adoption. Eight of the dogs and 6 of the cats are male. Find each of the following conditional probabilities if an animal is selected at random:
- The pet is male, given that it is a cat.
  - The pet is a cat, given that it is female.
  - The pet is female, given that it is a dog.
9. **Health.** The probabilities that an adult American man has high blood pressure and/or high cholesterol are shown in the table.

		Blood Pressure	
		High	OK
Cholesterol	High	0.11	0.21
	OK	0.16	0.52

What's the probability that

- a man has both conditions?
  - a man has high blood pressure?
  - a man with high blood pressure has high cholesterol?
  - a man has high blood pressure if it's known that he has high cholesterol?
10. **Death penalty.** The table shows the political affiliations of American voters and their positions on the death penalty.

		Death Penalty	
		Favor	Oppose
Party	Republican	0.26	0.04
	Democrat	0.12	0.24
	Other	0.24	0.10

- What's the probability that
  - a randomly chosen voter favors the death penalty?
  - a Republican favors the death penalty?
  - a voter who favors the death penalty is a Democrat?
- A candidate thinks she has a good chance of gaining the votes of anyone who is a Republican or in favor of the death penalty. What portion of the voters is that?

11. **Global survey, take 2.** Look again at the table summarizing the Roper survey in Exercise 5.
- If we select a respondent at random, what's the probability we choose a person from the United States who has done post-graduate study?
  - Among the respondents who have done post-graduate study, what's the probability the person is from the United States?
  - What's the probability that a respondent from the United States has done post-graduate study?
  - What's the probability that a respondent from China has only a primary-level education?
  - What's the probability that a respondent with only a primary-level education is from China?
12. **Birth order, take 2.** Look again at the data about birth order of Intro Stats students and their choices of colleges shown in Exercise 6.
- If we select a student at random, what's the probability the person is an Arts and Sciences student who is a second child (or more)?
  - Among the Arts and Sciences students, what's the probability a student was a second child (or more)?
  - Among second children (or more), what's the probability the student is enrolled in Arts and Sciences?
  - What's the probability that a first or only child is enrolled in the Agriculture College?
  - What is the probability that an Agriculture student is a first or only child?
13. **Sick kids.** Seventy percent of kids who visit a doctor have a fever, and 30% of kids with a fever have sore throats. What's the probability that a kid who goes to the doctor has a fever and a sore throat?
14. **Sick cars.** Twenty percent of cars that are inspected have faulty pollution control systems. The cost of repairing a pollution control system exceeds \$100 about 40% of the time. When a driver takes her car in for inspection, what's the probability that she will end up paying more than \$100 to repair the pollution control system?
15. **Cards.** You are dealt a hand of three cards, one at a time. Find the probability of each of the following.
- The first heart you get is the third card dealt.
  - Your cards are all red (that is, all diamonds or hearts).
  - You get no spades.
  - You have at least one ace.
16. **Another hand.** You pick three cards at random from a deck. Find the probability of each event described below.
- You get no aces.
  - You get all hearts.
  - The third card is your first red card.
  - You have at least one diamond.
17. **Batteries.** A junk box in your room contains a dozen old batteries, five of which are totally dead. You start picking batteries one at a time and testing them. Find the probability of each outcome.
- The first two you choose are both good.
  - At least one of the first three works.
  - The first four you pick all work.
  - You have to pick 5 batteries to find one that works.

18. **Shirts.** The soccer team's shirts have arrived in a big box, and people just start grabbing them, looking for the right size. The box contains 4 medium, 10 large, and 6 extra-large shirts. You want a medium for you and one for your sister. Find the probability of each event described.
- The first two you grab are the wrong sizes.
  - The first medium shirt you find is the third one you check.
  - The first four shirts you pick are all extra-large.
  - At least one of the first four shirts you check is a medium.
19. **Eligibility.** A university requires its biology majors to take a course called BioResearch. The prerequisite for this course is that students must have taken either a Statistics course or a computer course. By the time they are juniors, 52% of the Biology majors have taken Statistics, 23% have had a computer course, and 7% have done both.
- What percent of the junior Biology majors are ineligible for BioResearch?
  - What's the probability that a junior Biology major who has taken Statistics has also taken a computer course?
  - Are taking these two courses disjoint events? Explain.
  - Are taking these two courses independent events? Explain.
20. **Benefits.** Fifty-six percent of all American workers have a workplace retirement plan, 68% have health insurance, and 49% have both benefits. We select a worker at random.
- What's the probability he has neither employer-sponsored health insurance nor a retirement plan?
  - What's the probability he has health insurance if he has a retirement plan?
  - Are having health insurance and a retirement plan independent events? Explain.
  - Are having these two benefits mutually exclusive? Explain.
21. **For sale.** In the real-estate ads described in Exercise 1, 64% of homes for sale have garages, 21% have swimming pools, and 17% have both features.
- If a home for sale has a garage, what's the probability that it has a pool too?
  - Are having a garage and a pool independent events? Explain.
  - Are having a garage and a pool mutually exclusive? Explain.
22. **On the road again.** According to Exercise 2, the probability that a U.S. resident has traveled to Canada is 0.18, to Mexico is 0.09, and to both countries is 0.04.
- What's the probability that someone who has traveled to Mexico has visited Canada too?
  - Are traveling to Mexico and to Canada disjoint events? Explain.
  - Are traveling to Mexico and to Canada independent events? Explain.
23. **Cards.** If you draw a card at random from a well-shuffled deck, is getting an ace independent of the suit? Explain.
24. **Pets again.** The local animal shelter in Exercise 8 reported that it currently has 24 dogs and 18 cats available for adoption; 8 of the dogs and 6 of the cats are male. Are the species and sex of the animals independent? Explain.
25. **Unsafe food.** Early in 2007 *Consumer Reports* published the results of an extensive investigation of broiler chickens purchased from food stores in 23 states. Tests for bacteria in the meat showed that 81% of the chickens were contaminated with campylobacter, 15% with salmonella, and 13% with both.
- What's the probability that a tested chicken was not contaminated with either kind of bacteria?
  - Are contamination with the two kinds of bacteria disjoint? Explain.
  - Are contamination with the two kinds of bacteria independent? Explain.
26. **Birth order, finis.** In Exercises 6 and 12 we looked at the birth orders and college choices of some Intro Stats students. For these students:
- Are enrolling in Agriculture and Human Ecology disjoint? Explain.
  - Are enrolling in Agriculture and Human Ecology independent? Explain.
  - Are being firstborn and enrolling in Human Ecology disjoint? Explain.
  - Are being firstborn and enrolling in Human Ecology independent? Explain.
27. **Men's health, again.** Given the table of probabilities from Exercise 9, are high blood pressure and high cholesterol independent? Explain.

		Blood Pressure	
		High	OK
Cholesterol	High	0.11	0.21
	OK	0.16	0.52

28. **Politics.** Given the table of probabilities from Exercise 10, are party affiliation and position on the death penalty independent? Explain.

		Death Penalty	
		Favor	Oppose
Party	Republican	0.26	0.04
	Democrat	0.12	0.24
	Other	0.24	0.10

29. **Phone service.** According to estimates from the federal government's 2003 National Health Interview Survey, based on face-to-face interviews in 16,677 households, approximately 58.2% of U.S. adults have both a landline in their residence and a cell phone, 2.8% have only cell phone service but no landline, and 1.6% have no telephone service at all.

- a) Polling agencies won't phone cell phone numbers because customers object to paying for such calls. What proportion of U.S. households can be reached by a landline call?
- b) Are having a cell phone and having a landline independent? Explain.
30. **Snoring.** After surveying 995 adults, 81.5% of whom were over 30, the National Sleep Foundation reported that 36.8% of all the adults snored. 32% of the respondents were snorers over the age of 30.
- a) What percent of the respondents were under 30 and did not snore?
- b) Is snoring independent of age? Explain.
31. **Montana.** A 1992 poll conducted by the University of Montana classified respondents by sex and political party, as shown in the table. Is party affiliation independent of the respondents' sex? Explain.

	Democrat	Republican	Independent
Male	36	45	24
Female	48	33	16

32. **Cars.** A random survey of autos parked in student and staff lots at a large university classified the brands by country of origin, as seen in the table. Is country of origin independent of type of driver?

		Driver	
		Student	Staff
Origin	American	107	105
	European	33	12
	Asian	55	47

33. **Luggage.** Leah is flying from Boston to Denver with a connection in Chicago. The probability her first flight leaves on time is 0.15. If the flight is on time, the probability that her luggage will make the connecting flight in Chicago is 0.95, but if the first flight is delayed, the probability that the luggage will make it is only 0.65.
- a) Are the first flight leaving on time and the luggage making the connection independent events? Explain.
- b) What is the probability that her luggage arrives in Denver with her?
34. **Graduation.** A private college report contains these statistics:
- 70% of incoming freshmen attended public schools.*  
*75% of public school students who enroll as freshmen eventually graduate.*  
*90% of other freshmen eventually graduate.*
- a) Is there any evidence that a freshman's chances to graduate may depend upon what kind of high school the student attended? Explain.
- b) What percent of freshmen eventually graduate?
35. **Late luggage.** Remember Leah (Exercise 33)? Suppose you pick her up at the Denver airport, and her luggage is

not there. What is the probability that Leah's first flight was delayed?

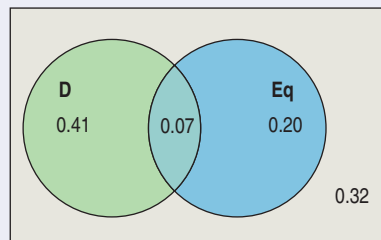
36. **Graduation, part II.** What percent of students who graduate from the college in Exercise 34 attended a public high school?
37. **Absenteeism.** A company's records indicate that on any given day about 1% of their day-shift employees and 2% of the night-shift employees will miss work. Sixty percent of the employees work the day shift.
- a) Is absenteeism independent of shift worked? Explain.
- b) What percent of employees are absent on any given day?
38. **Lungs and smoke.** Suppose that 23% of adults smoke cigarettes. It's known that 57% of smokers and 13% of nonsmokers develop a certain lung condition by age 60.
- a) Explain how these statistics indicate that lung condition and smoking are not independent.
- b) What's the probability that a randomly selected 60-year-old has this lung condition?
39. **Absenteeism, part II.** At the company described in Exercise 37, what percent of the absent employees are on the night shift?
40. **Lungs and smoke again.** Based on the statistics in Exercise 38, what's the probability that someone with the lung condition was a smoker?
41. **Drunks.** Police often set up sobriety checkpoints—roadblocks where drivers are asked a few brief questions to allow the officer to judge whether or not the person may have been drinking. If the officer does not suspect a problem, drivers are released to go on their way. Otherwise, drivers are detained for a Breathalyzer test that will determine whether or not they will be arrested. The police say that based on the brief initial stop, trained officers can make the right decision 80% of the time. Suppose the police operate a sobriety checkpoint after 9 p.m. on a Saturday night, a time when national traffic safety experts suspect that about 12% of drivers have been drinking.
- a) You are stopped at the checkpoint and, of course, have not been drinking. What's the probability that you are detained for further testing?
- b) What's the probability that any given driver will be detained?
- c) What's the probability that a driver who is detained has actually been drinking?
- d) What's the probability that a driver who was released had actually been drinking?
42. **No-shows.** An airline offers discounted "advance-purchase" fares to customers who buy tickets more than 30 days before travel and charges "regular" fares for tickets purchased during those last 30 days. The company has noticed that 60% of its customers take advantage of the advance-purchase fares. The "no-show" rate among people who paid regular fares is 30%, but only 5% of customers with advance-purchase tickets are no-shows.
- a) What percent of all ticket holders are no-shows?
- b) What's the probability that a customer who didn't show had an advance-purchase ticket?
- c) Is being a no-show independent of the type of ticket a passenger holds? Explain.

43. **Dishwashers.** Dan's Diner employs three dishwashers. Al washes 40% of the dishes and breaks only 1% of those he handles. Betty and Chuck each wash 30% of the dishes, and Betty breaks only 1% of hers, but Chuck breaks 3% of the dishes he washes. (He, of course, will need a new job soon. . . .) You go to Dan's for supper one night and hear a dish break at the sink. What's the probability that Chuck is on the job?
44. **Parts.** A company manufacturing electronic components for home entertainment systems buys electrical connectors from three suppliers. The company prefers to use supplier A because only 1% of those connectors prove to be defective, but supplier A can deliver only 70% of the connectors needed. The company must also purchase connectors from two other suppliers, 20% from supplier B and the rest from supplier C. The rates of defective connectors from B and C are 2% and 4%, respectively. You buy one of these components, and when you try to use it you find that the connector is defective. What's the probability that your component came from supplier A?
45. **HIV testing.** In July 2005 the journal *Annals of Internal Medicine* published a report on the reliability of HIV testing. Results of a large study suggested that among people with HIV, 99.7% of tests conducted were (correctly) positive, while for people without HIV 98.5% of the tests were (correctly) negative. A clinic serving an at-risk population offers free HIV testing, believing that 15% of the patients may actually carry HIV. What's the probability that a patient testing negative is truly free of HIV?
46. **Polygraphs.** Lie detectors are controversial instruments, barred from use as evidence in many courts. Nonetheless, many employers use lie detector screening as part of their hiring process in the hope that they can avoid hiring people who might be dishonest. There has been some research, but no agreement, about the reliability of polygraph tests. Based on this research, suppose that a polygraph can detect 65% of lies, but incorrectly identifies 15% of true statements as lies.
- A certain company believes that 95% of its job applicants are trustworthy. The company gives everyone a polygraph test, asking, "Have you ever stolen anything from your place of work?" Naturally, all the applicants answer "No," but the polygraph identifies some of those answers as lies, making the person ineligible for a job. What's the probability that a job applicant rejected under suspicion of dishonesty was actually trustworthy?



## JUST CHECKING Answers

1. a)



b) 0.32

c) 0.41

2. a) Independent

b) Disjoint

c) Neither

3. a)

		Equation		
		Yes	No	Total
Display	Yes	0.07	0.41	<b>0.48</b>
	No	0.20	0.32	<b>0.52</b>
	Total	<b>0.27</b>	<b>0.73</b>	<b>1.00</b>

b)  $P(D | Eq) = P(D \text{ and } Eq) / P(Eq) = 0.07 / 0.27 = 0.259$

c) No, pages can (and 7% do) have both.

d) To be independent, we'd need  $P(D | Eq) = P(D)$ .  $P(D | Eq) = 0.259$ , but  $P(D) = 0.48$ . Overall, 48% of pages have data displays, but only about 26% of pages with equations do. They do not appear to be independent.

# Random Variables



## What Is an Actuary?

Actuaries are the daring people who put a price on risk, estimating the likelihood and costs of rare events, so they can be insured. That takes financial, statistical, and business skills. It also makes them invaluable to many businesses. Actuaries are rather rare themselves; only about 19,000 work in North America. Perhaps because of this, they are well paid. If you're enjoying this course, you may want to look into a career as an actuary. Contact the Society of Actuaries or the Casualty Actuarial Society (who, despite what you may think, did not pay for this blurb).

Insurance companies make bets. They bet that you're going to live a long life. You bet that you're going to die sooner. Both you and the insurance company want the company to stay in business, so it's important to find a "fair price" for your bet. Of course, the right price for *you* depends on many factors, and nobody can predict exactly how long you'll live. But when the company averages over enough customers, it can make reasonably accurate estimates of the amount it can expect to collect on a policy before it has to pay its benefit.

Here's a simple example. An insurance company offers a "death and disability" policy that pays \$10,000 when you die or \$5000 if you are permanently disabled. It charges a premium of only \$50 a year for this benefit. Is the company likely to make a profit selling such a plan? To answer this question, the company needs to know the *probability* that its clients will die or be disabled in any year. From actuarial information like this, the company can calculate the expected value of this policy.

## Expected Value: Center

### NOTATION ALERT:

The most common letters for random variables are  $X$ ,  $Y$ , and  $Z$ . But be cautious: If you see any capital letter, it just might denote a random variable.

We'll want to build a probability model in order to answer the questions about the insurance company's risk. First we need to define a few terms. The amount the company pays out on an individual policy is called a **random variable** because its numeric value is based on the outcome of a random event. We use a capital letter, like  $X$ , to denote a random variable. We'll denote a particular value that it can have by the corresponding lowercase letter, in this case  $x$ . For the insurance company,  $x$  can be \$10,000 (if you die that year), \$5000 (if you are disabled), or \$0 (if neither occurs). Because we can list all the outcomes, we might formally call this random variable a **discrete** random variable. Otherwise, we'd call it a **continuous** random variable. The collection of all the possible values and the probabilities that they occur is called the **probability model** for the random variable.

**AS** **Activity: Random Variables.** Learn more about random variables from this animated tour.

Suppose, for example, that the death rate in any year is 1 out of every 1000 people, and that another 2 out of 1000 suffer some kind of disability. Then we can display the probability model for this insurance policy in a table like this:

Policyholder Outcome	Payout $x$	Probability $P(X = x)$
Death	10,000	$\frac{1}{1000}$
Disability	5000	$\frac{2}{1000}$
Neither	0	$\frac{997}{1000}$

To see what the insurance company can expect, imagine that it insures exactly 1000 people. Further imagine that, in perfect accordance with the probabilities, 1 of the policyholders dies, 2 are disabled, and the remaining 997 survive the year unscathed. The company would pay \$10,000 to one client and \$5000 to each of 2 clients. That's a total of \$20,000, or an average of  $20000/1000 = \$20$  per policy. Since it is charging people \$50 for the policy, the company expects to make a profit of \$30 per customer. Not bad!

#### NOTATION ALERT:

The expected value (or mean) of a random variable is written  $E(X)$  or  $\mu$ .

We can't predict what *will* happen during any given year, but we can say what we *expect* to happen. To do this, we (or, rather, the insurance company) need the probability model. The expected value of a policy is a parameter of this model. In fact, it's the mean. We'll signify this with the notation  $\mu$  (for population mean) or  $E(X)$  for expected value. This isn't an average of some data values, so we won't estimate it. Instead, we assume that the probabilities are known and simply calculate the expected value from them.

How did we come up with \$20 as the expected value of a policy payout? Here's the calculation. As we've seen, it often simplifies probability calculations to think about some (convenient) number of outcomes. For example, we could imagine that we have exactly 1000 clients. Of those, exactly 1 died and 2 were disabled, corresponding to what the probabilities would say.

$$\mu = E(X) = \frac{10,000(1) + 5000(2) + 0(997)}{1000}$$

So our expected payout comes to \$20,000, or \$20 per policy.

Instead of writing the expected value as one big fraction, we can rewrite it as separate terms with a common denominator of 1000.

$$\begin{aligned}\mu &= E(X) \\ &= \$10,000\left(\frac{1}{1000}\right) + \$5000\left(\frac{2}{1000}\right) + \$0\left(\frac{997}{1000}\right) \\ &= \$20.\end{aligned}$$

How convenient! See the probabilities? For each policy, there's a  $1/1000$  chance that we'll have to pay \$10,000 for a death and a  $2/1000$  chance that we'll have to pay \$5000 for a disability. Of course, there's a  $997/1000$  chance that we won't have to pay anything.

Take a good look at the expression now. It's easy to calculate the **expected value** of a (discrete) random variable—just multiply each possible value by the probability that it occurs, and find the sum:

$$\mu = E(X) = \sum xP(x).$$

Be sure that every possible outcome is included in the sum. And verify that you have a valid probability model to start with—the probabilities should each be between 0 and 1 and should sum to one.

## FOR EXAMPLE

## Love and expected values

On Valentine's Day the *Quiet Nook* restaurant offers a *Lucky Lovers Special* that could save couples money on their romantic dinners. When the waiter brings the check, he'll also bring the four aces from a deck of cards. He'll shuffle them and lay them out face down on the table. The couple will then get to turn one card over. If it's a black ace, they'll owe the full amount, but if it's the ace of hearts, the waiter will give them a \$20 Lucky Lovers discount. If they first turn over the ace of diamonds (hey—at least it's red!), they'll then get to turn over one of the remaining cards, earning a \$10 discount for finding the ace of hearts this time.

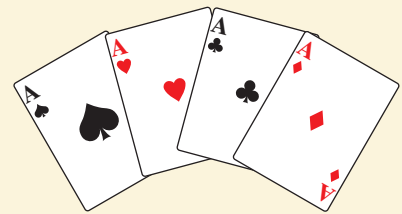
**Question:** Based on a probability model for the size of the Lucky Lovers discounts the restaurant will award, what's the expected discount for a couple?

Let  $X$  = the Lucky Lovers discount. The probabilities of the three outcomes are:

$$P(X = 20) = P(A \heartsuit) = \frac{1}{4}$$

$$\begin{aligned} P(X = 10) &= P(A \spadesuit, \text{ then } A \heartsuit) = P(A \spadesuit) \cdot P(A \heartsuit | A \spadesuit) \\ &= \frac{1}{4} \cdot \frac{1}{3} = \frac{1}{12} \end{aligned}$$

$$P(X = 0) = P(X \neq 20 \text{ or } 10) = 1 - \left( \frac{1}{4} + \frac{1}{12} \right) = \frac{2}{3}$$



My probability model is:

Outcome	A $\heartsuit$	A $\spadesuit$ , then A $\heartsuit$	Black Ace
$x$	20	10	0
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{2}{3}$

$$E(X) = 20 \cdot \frac{1}{4} + 10 \cdot \frac{1}{12} + 0 \cdot \frac{2}{3} = \frac{70}{12} \approx 5.83$$

Couples dining at the Quiet Nook can expect an average discount of \$5.83.



## JUST CHECKING

- One of the authors took his minivan in for repair recently because the air conditioner was cutting out intermittently. The mechanic identified the problem as dirt in a control unit. He said that in about 75% of such cases, drawing down and then recharging the coolant a couple of times cleans up the problem—and costs only \$60. If that fails, then the control unit must be replaced at an additional cost of \$100 for parts and \$40 for labor.
  - Define the random variable and construct the probability model.
  - What is the expected value of the cost of this repair?
  - What does that mean in this context?

Oh—in case you were wondering—the \$60 fix worked!

## First Center, Now Spread . . .

Of course, this expected value (or mean) is not what actually happens to any *particular* policyholder. No individual policy actually costs the company \$20. We are dealing with random events, so some policyholders receive big payouts, others nothing. Because the insurance company must anticipate this variability, it needs to know the *standard deviation* of the random variable.

For data, we calculated the **standard deviation** by first computing the deviation from the mean and squaring it. We do that with (discrete) random variables as well. First, we find the deviation of each payout from the mean (expected value):

Policyholder Outcome	Payout $x$	Probability $P(X = x)$	Deviation $(x - \mu)$
Death	10,000	$\frac{1}{1000}$	$(10,000 - 20) = 9980$
Disability	5000	$\frac{2}{1000}$	$(5000 - 20) = 4980$
Neither	0	$\frac{997}{1000}$	$(0 - 20) = -20$

Next, we square each deviation. **The variance** is the expected value of those squared deviations, so we multiply each by the appropriate probability and sum those products. That gives us the variance of  $X$ . Here's what it looks like:

$$\text{Var}(X) = 9980^2 \left( \frac{1}{1000} \right) + 4980^2 \left( \frac{2}{1000} \right) + (-20)^2 \left( \frac{997}{1000} \right) = 149,600.$$

Finally, we take the square root to get the standard deviation:

$$\text{SD}(X) = \sqrt{149,600} \approx \$386.78.$$

The insurance company can expect an average payout of \$20 per policy, with a standard deviation of \$386.78.

Think about that. The company charges \$50 for each policy and expects to pay out \$20 per policy. Sounds like an easy way to make \$30. In fact, most of the time (probability 997/1000) the company pockets the entire \$50. But would you consider selling your roommate such a policy? The problem is that occasionally the company loses big. With probability 1/1000, it will pay out \$10,000, and with probability 2/1000, it will pay out \$5000. That may be more risk than you're willing to take on. The standard deviation of \$386.78 gives an indication that it's no sure thing. That's a pretty big spread (and risk) for an average profit of \$30.

Here are the formulas for what we just did. Because these are parameters of our probability model, the variance and standard deviation can also be written as  $\sigma^2$  and  $\sigma$ . You should recognize both kinds of notation.

$$\begin{aligned} \sigma^2 &= \text{Var}(X) = \sum (x - \mu)^2 P(x) \\ \sigma &= \text{SD}(X) = \sqrt{\text{Var}(X)} \end{aligned}$$



## FOR EXAMPLE

## Finding the standard deviation

**Recap:** Here's the probability model for the Lucky Lovers restaurant discount.

Outcome	A♥	A♠, then A♥	Black Ace
$x$	20	10	0
$P(X = x)$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{2}{3}$

We found that couples can expect an average discount of  $\mu = \$5.83$ .

**Question:** What's the standard deviation of the discounts?

First find the variance:  $\text{Var}(X) = \sum (x - \mu)^2 \cdot P(x)$

$$\begin{aligned}
 &= (20 - 5.83)^2 \cdot \frac{1}{4} + (10 - 5.83)^2 \cdot \frac{1}{12} + (0 - 5.83)^2 \cdot \frac{2}{3} \\
 &\approx 74.306.
 \end{aligned}$$

So,  $SD(X) = \sqrt{74.306} \approx \$8.62$

Couples can expect the Lucky Lovers discounts to average \$5.83, with a standard deviation of \$8.62.

## STEP-BY-STEP EXAMPLE

## Expected Values and Standard Deviations for Discrete Random Variables

As the head of inventory for Knowway computer company, you were thrilled that you had managed to ship 2 computers to your biggest client the day the order arrived. You are horrified, though, to find out that someone had restocked refurbished computers in with the new computers in your storeroom. The shipped computers were selected randomly from the 15 computers in stock, but 4 of those were actually refurbished.

If your client gets 2 new computers, things are fine. If the client gets one refurbished computer, it will be sent back at your expense—\$100—and you can replace it. However, if both computers are refurbished, the client will cancel the order this month and you'll lose a total of \$1000.

**Question:** What's the expected value and the standard deviation of the company's loss?



**Plan** State the problem.

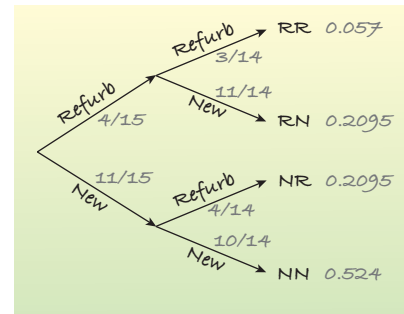
**Variable** Define the random variable.

I want to find the company's expected loss for shipping refurbished computers and the standard deviation.

Let  $X =$  amount of loss.

**Plot** Make a picture. This is another job for tree diagrams.

If you prefer calculation to drawing, find  $P(\text{NN})$  and  $P(\text{RR})$ , then use the Complement Rule to find  $P(\text{NR or RN})$ .



**Model** List the possible values of the random variable, and determine the probability model.

Outcome	x	P(X = x)
Two refurb	1000	$P(\text{RR}) = 0.057$
One refurb	100	$P(\text{NR} \cup \text{RN}) = 0.2095 + 0.2095 = 0.419$
New/new	0	$P(\text{NN}) = 0.524$



**Mechanics** Find the expected value.

Find the variance.

Find the standard deviation.

$$E(X) = 0(0.524) + 100(0.419) + 1000(0.057) = \$98.90$$

$$\begin{aligned} \text{Var}(X) &= (0 - 98.90)^2(0.524) \\ &\quad + (100 - 98.90)^2(0.419) \\ &\quad + (1000 - 98.90)^2(0.057) \\ &= 51,408.79 \end{aligned}$$

$$SD(X) = \sqrt{51,408.79} = \$226.735$$



**Conclusion** Interpret your results in context.



Both numbers seem reasonable. The expected value of \$98.90 is between the extremes of \$0 and \$1000, and there's great variability in the outcome values.

I expect this mistake to cost the firm \$98.90, with a standard deviation of \$226.74. The large standard deviation reflects the fact that there's a pretty large range of possible losses.

**TI Tips**

**Finding the mean and SD of a random variable**

L1	L2	L3	Z
0	.52381		----
100	.41905		
1000			
----			
L2(3) = 4/15 * 3/14			

You can easily calculate means and standard deviations for a random variable with your TI. Let's do the Knowway computer example.

- Enter the values of the variable in a list, say, **L1**: 0, 100, 1000.
- Enter the probability model in another list, say, **L2**. Notice that you can enter the probabilities as fractions. For example, multiplying along the top branches

```
1-Var Stats L1,L2
Σ
```

```
1-Var Stats
x̄=99.04761905
Σx=99.04761905
Σx²=61333.3333
Sx=
σx=226.986569
↓n=1
```

of the tree gives the probability of a \$1000 loss to be  $\frac{4}{15} \cdot \frac{3}{14}$ . When you enter that, the TI will automatically calculate the probability as a decimal!

- Under the **STAT CALC** menu, ask for **1-Var Stats L1,L2**.

Now you see the mean and standard deviation (along with some other things). Don't fret that the calculator's mean and standard deviation aren't precisely the same as the ones we found. Such minor differences can arise whenever we round off probabilities to do the work by hand.

Beware: Although the calculator knows enough to call the standard deviation  $\sigma$ , it uses  $\bar{x}$  where it should say  $\mu$ . Make sure you don't make that mistake!

## More About Means and Variances

Our insurance company expected to pay out an average of \$20 per policy, with a standard deviation of about \$387. If we take the \$50 premium into account, we see the company makes a profit of  $50 - 20 = \$30$  per policy. Suppose the company lowers the premium by \$5 to \$45. It's pretty clear that the expected profit also drops an average of \$5 per policy, to  $45 - 20 = \$25$ .

What about the standard deviation? We know that adding or subtracting a constant from data shifts the mean but doesn't change the variance or standard deviation. The same is true of random variables.<sup>1</sup>

$$E(X \pm c) = E(X) \pm c \quad \text{Var}(X \pm c) = \text{Var}(X).$$

### FOR EXAMPLE

#### Adding a constant

**Recap:** We've determined that couples dining at the *Quiet Nook* can expect Lucky Lovers discounts averaging \$5.83 with a standard deviation of \$8.62. Suppose that for several weeks the restaurant has also been distributing coupons worth \$5 off any one meal (one discount per table).

**Question:** If every couple dining there on Valentine's Day brings a coupon, what will be the mean and standard deviation of the total discounts they'll receive?

Let  $D$  = total discount (Lucky Lovers plus the coupon); then  $D = X + 5$ .

$$\begin{aligned} E(D) &= E(X + 5) = E(X) + 5 = 5.83 + 5 = \$10.83 \\ \text{Var}(D) &= \text{Var}(X + 5) = \text{Var}(X) = 8.62^2 \\ \text{SD}(D) &= \sqrt{\text{Var}(X)} = \$8.62 \end{aligned}$$

Couples with the coupon can expect total discounts averaging \$10.83. The standard deviation is still \$8.62.

Back to insurance . . . What if the company decides to double all the payouts—that is, pay \$20,000 for death and \$10,000 for disability? This would double the average payout per policy and also increase the variability in payouts. We have seen that multiplying or dividing all data values by a constant changes both the mean and the standard deviation by the same factor. Variance, being the square of standard deviation, changes by the square of the constant. The same is true of random variables. In general, multiplying each value of a random variable by a

<sup>1</sup> The rules in this section are true for both discrete and continuous random variables.

constant multiplies the mean by that constant and the variance by the *square* of the constant.

$$E(aX) = aE(X) \quad \text{Var}(aX) = a^2\text{Var}(X)$$

### FOR EXAMPLE

#### Double the love

**Recap:** On Valentine's Day at the *Quiet Nook*, couples may get a Lucky Lovers discount averaging \$5.83 with a standard deviation of \$8.62. When two couples dine together on a single check, the restaurant doubles the discount offer—\$40 for the ace of hearts on the first card and \$20 on the second.

**Question:** What are the mean and standard deviation of discounts for such foursomes?

$$E(2X) = 2E(X) = 2(5.83) = \$11.66$$

$$\text{Var}(2x) = 2^2\text{Var}(x) = 2^2 \cdot 8.62^2 = 297.2176$$

$$\text{SD}(2X) = \sqrt{297.2176} = \$17.24$$

If the restaurant doubles the discount offer, two couples dining together can expect to save an average of \$11.66 with a standard deviation of \$17.24.

This insurance company sells policies to more than just one person. How can we figure means and variances for a collection of customers? For example, how can the company find the total expected value (and standard deviation) of policies taken over all policyholders? Consider a simple case: just two customers, Mr. Ecks and Ms. Wye. With an expected payout of \$20 on each policy, we might predict a total of  $\$20 + \$20 = \$40$  to be paid out on the two policies. Nothing surprising there. The expected value of the sum is the sum of the expected values.

$$E(X + Y) = E(X) + E(Y).$$

The variability is another matter. Is the risk of insuring two people the same as the risk of insuring one person for twice as much? We wouldn't expect both clients to die or become disabled in the same year. Because we've spread the risk, the standard deviation should be smaller. Indeed, this is the fundamental principle behind insurance. By spreading the risk among many policies, a company can keep the standard deviation quite small and predict costs more accurately.

But how much smaller is the standard deviation of the sum? It turns out that, if the random variables are independent, there is a simple Addition Rule for variances: *The variance of the sum of two independent random variables is the sum of their individual variances.*

For Mr. Ecks and Ms. Wye, the insurance company can expect their outcomes to be independent, so (using  $X$  for Mr. Ecks's payout and  $Y$  for Ms. Wye's)

$$\begin{aligned} \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) \\ &= 149,600 + 149,600 \\ &= 299,200. \end{aligned}$$

If they had insured only Mr. Ecks for twice as much, there would only be one outcome rather than two *independent* outcomes, so the variance would have been

$$\text{Var}(2X) = 2^2\text{Var}(X) = 4 \times 149,600 = 598,400, \text{ or}$$

twice as big as with two independent policies.

Of course, variances are in squared units. The company would prefer to know standard deviations, which are in dollars. The standard deviation of the payout for two independent policies is  $\sqrt{299,200} = \$546.99$ . But the standard deviation

of the payout for a single policy of twice the size is  $\sqrt{598,400} = \$773.56$ , or about 40% more.

If the company has two customers, then, it will have an expected annual total payout of \$40 with a standard deviation of about \$547.

## FOR EXAMPLE

### Adding the discounts

**Recap:** The Valentine's Day Lucky Lovers discount for couples averages \$5.83 with a standard deviation of \$8.62. We've seen that if the restaurant doubles the discount offer for two couples dining together on a single check, they can expect to save \$11.66 with a standard deviation of \$17.24. Some couples decide instead to get separate checks and pool their two discounts.

**Question:** You and your amour go to this restaurant with another couple and agree to share any benefit from this promotion. Does it matter whether you pay separately or together?

Let  $X_1$  and  $X_2$  represent the two separate discounts, and  $T$  the total; then  $T = X_1 + X_2$ .

$$E(T) = E(X_1 + X_2) = E(X_1) + E(X_2) = 5.83 + 5.83 = \$11.66,$$

so the expected saving is the same either way.

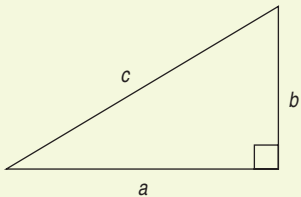
The cards are reshuffled for each couple's turn, so the discounts couples receive are independent. It's okay to add the variances:

$$\begin{aligned} \text{Var}(T) &= \text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = 8.62^2 + 8.62^2 = 148.6088 \\ \text{SD}(T) &= \sqrt{148.6088} = \$12.19 \end{aligned}$$

When two couples get separate checks, there's less variation in their total discount. The standard deviation is \$12.19, compared to \$17.24 for couples who play for the double discount on a single check. It does, therefore, matter whether they pay separately or together.

### Pythagorean Theorem of Statistics

We often use the standard deviation to measure variability, but when we add independent random variables, we use their variances. Think of the Pythagorean Theorem. In a right triangle (only), the square of the length of the hypotenuse is the sum of the squares of the lengths of the other two sides:



$$c^2 = a^2 + b^2.$$

For independent random variables (only), the square of the standard deviation of their sum is the sum of the squares of their standard deviations:

$$\text{SD}^2(X + Y) = \text{SD}^2(X) + \text{SD}^2(Y).$$

It's simpler to write this with variances:

For independent random variables,  $X$  and  $Y$ ,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

In general,

- ▶ The mean of the sum of two random variables is the sum of the means.
- ▶ The mean of the difference of two random variables is the difference of the means.
- ▶ If the random variables are independent, the variance of their sum or difference is always the sum of the variances.

$$E(X \pm Y) = E(X) \pm E(Y) \quad \text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$$

Wait a minute! Is that third part correct? Do we always *add* variances? Yes. Think about the two insurance policies. Suppose we want to know the mean and standard deviation of the *difference* in payouts to the two clients. Since each policy has an expected payout of \$20, the expected difference is  $20 - 20 = \$0$ . If we also subtract variances, we get \$0, too, and that surely doesn't make sense. Note that if the outcomes for the two clients are independent, the difference in payouts could range from  $\$10,000 - \$0 = \$10,000$  to  $\$0 - \$10,000 = -\$10,000$ , a spread of \$20,000. The variability in differences increases as much as the variability in sums. If the company has two customers, the difference in payouts has a mean of \$0 and a standard deviation of about \$547 (again).

## FOR EXAMPLE

## Working with differences

**Recap:** The Lucky Lovers discount at the *Quiet Nook* averages \$5.83 with a standard deviation of \$8.62. Just up the street, the *Wise Fool* restaurant has a competing Lottery of Love promotion. There a couple can select a specially prepared chocolate from a large bowl and unwrap it to learn the size of their discount. The restaurant's manager says the discounts vary with an average of \$10.00 and a standard deviation of \$15.00.

**Question:** How much more can you expect to save at the *Wise Fool*? With what standard deviation?

Let  $W$  = discount at the *Wise Fool*,  $X$  = the discount at the *Quiet Nook*, and  $D$  = the difference:  $D = W - X$ . These are different promotions at separate restaurants, so the outcomes are independent.

$$\begin{aligned} E(W - X) &= E(W) - E(X) = 10.00 - 5.83 = \$4.17 \\ SD(W - X) &= \sqrt{\text{Var}(W - X)} \\ &= \sqrt{\text{Var}(W) + \text{Var}(X)} \\ &= \sqrt{15^2 + 8.62^2} \\ &\approx \$17.30 \end{aligned}$$

Discounts at the *Wise Fool* will average \$4.17 more than at the *Quiet Nook*, with a standard deviation of \$17.30.

**For random variables, does  $X + X + X = 3X$ ?** Maybe, but be careful. As we've just seen, insuring one person for \$30,000 is not the same risk as insuring three people for \$10,000 each. When each instance represents a different outcome for the same random variable, it's easy to fall into the trap of writing all of them with the same symbol. Don't make this common mistake. Make sure you write each instance as a *different* random variable. Just because each random variable describes a similar situation doesn't mean that each random outcome will be the same.

These are *random* variables, not the variables you saw in Algebra. Being random, they take on different values each time they're evaluated. So what you really mean is  $X_1 + X_2 + X_3$ . Written this way, it's clear that the sum shouldn't necessarily equal 3 times *anything*.

## FOR EXAMPLE

## Summing a series of outcomes

**Recap:** The *Quiet Nook's* Lucky Lovers promotion offers couples discounts averaging \$5.83 with a standard deviation of \$8.62. The restaurant owner is planning to serve 40 couples on Valentine's Day.

**Question:** What's the expected total of the discounts the owner will give? With what standard deviation?

Let  $X_1, X_2, X_3, \dots, X_{40}$  represent the discounts to the 40 couples, and  $T$  the total of all the discounts. Then:

$$\begin{aligned} T &= X_1 + X_2 + X_3 + \dots + X_{40} \\ E(T) &= E(X_1 + X_2 + X_3 + \dots + X_{40}) \\ &= E(X_1) + E(X_2) + E(X_3) + \dots + E(X_{40}) \\ &= 5.83 + 5.83 + 5.83 + \dots + 5.83 \\ &= \$233.20 \end{aligned}$$

Reshuffling cards between couples makes the discounts independent, so:

$$\begin{aligned} SD(T) &= \sqrt{\text{Var}(X_1 + X_2 + X_3 + \dots + X_{40})} \\ &= \sqrt{\text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \dots + \text{Var}(X_{40})} \\ &= \sqrt{8.62^2 + 8.62^2 + 8.62^2 + \dots + 8.62^2} \\ &\approx \$54.52 \end{aligned}$$

The restaurant owner can expect the 40 couples to win discounts totaling \$233.20, with a standard deviation of \$54.52.



## JUST CHECKING

2. Suppose the time it takes a customer to get and pay for seats at the ticket window of a baseball park is a random variable with a mean of 100 seconds and a standard deviation of 50 seconds. When you get there, you find only two people in line in front of you.
- How long do you expect to wait for your turn to get tickets?
  - What's the standard deviation of your wait time?
  - What assumption did you make about the two customers in finding the standard deviation?

### STEP-BY-STEP EXAMPLE

### Hitting the Road: Means and Variances

You're planning to spend next year wandering through the mountains of Kyrgyzstan. You plan to sell your used SUV so you can purchase an off-road Honda motor scooter when you get there. Used SUVs of the year and mileage of yours are selling for a mean of \$6940 with a standard deviation of \$250. Your research shows that scooters in Kyrgyzstan are going for about 65,000 Kyrgyzstan som with a standard deviation of 500 som. One U.S. dollar is worth about 38.5 Kyrgyzstan som (38 som and 50 tlyn).

**Question:** How much cash can you expect to pocket after you sell your SUV and buy the scooter?



**Plan** State the problem.

**Variables** Define the random variables.

Write an appropriate equation.  
Think about the assumptions.

I want to model how much money I'd have (in som) after selling my SUV and buying the scooter.

Let  $A$  = sale price of my SUV (in dollars),  
 $B$  = price of a scooter (in som), and  
 $D$  = profit (in som)

$$D = 38.5A - B$$

✓ **Independence Assumption:** The prices are independent.



**Mechanics** Find the expected value, using the appropriate rules.

$$\begin{aligned} E(D) &= E(38.5A - B) \\ &= 38.5E(A) - E(B) \\ &= 38.5(6,940) - (65,000) \\ E(D) &= 202,190 \text{ som} \end{aligned}$$

Find the variance, using the appropriate rules. Be sure to check the assumptions first!

Since sale and purchase prices are independent,

$$\begin{aligned} \text{Var}(D) &= \text{Var}(38.5A - B) \\ &= \text{Var}(38.5A) + \text{Var}(B) \\ &= (38.5)^2 \text{Var}(A) + \text{Var}(B) \\ &= 1482.25(250)^2 + (500)^2 \\ \text{Var}(D) &= 92,890,625 \end{aligned}$$

Find the standard deviation.

$$SD(D) = \sqrt{92,890,625} = 9637.98 \text{ som}$$



**Conclusion** Interpret your results in context. (Here that means talking about dollars.)

I can expect to clear about 202,190 som (\$5252) with a standard deviation of 9638 som (\$250).



Given the initial cost estimates, the mean and standard deviation seem reasonable.

## Continuous Random Variables

**A S**

**Activity: Numeric**

**Outcomes.** You've seen how to simulate discrete random outcomes. There's a tool for simulating continuous outcomes, too.

**A S**

**Activity: Means of Random Variables.** Experiment with continuous random variables to learn how their expected values behave.

A company manufactures small stereo systems. At the end of the production line, the stereos are packaged and prepared for shipping. Stage 1 of this process is called “packing.” Workers must collect all the system components (a main unit, two speakers, a power cord, an antenna, and some wires), put each in plastic bags, and then place everything inside a protective styrofoam form. The packed form then moves on to Stage 2, called “boxing.” There, workers place the form and a packet of instructions in a cardboard box, close it, then seal and label the box for shipping.

The company says that times required for the packing stage can be described by a Normal model with a mean of 9 minutes and standard deviation of 1.5 minutes. The times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute.

This is a common way to model events. Do our rules for random variables apply here? What's different? We no longer have a list of discrete outcomes, with their associated probabilities. Instead, we have **continuous random variables that can take on any value**. Now any single value won't have a probability. We saw this back in Chapter 6 when we first saw the Normal model (although we didn't talk then about “random variables” or “probability”). We know that the probability that  $z = 1.5$  doesn't make sense, but we *can* talk about the probability that  $z$  lies *between* 0.5 and 1.5. For a Normal random variable, the probability that it falls within an interval is just the area under the Normal curve over that interval.

Some continuous random variables have Normal models; others may be skewed, uniform, or bimodal. Regardless of shape, all continuous random variables have means (which we also call *expected values*) and variances. In this book we won't worry about how to calculate them, but we can still work with models for continuous random variables when we're given these parameters.

The good news is that nearly everything we've said about how discrete random variables behave is true of continuous random variables, as well. **When two independent continuous random variables have Normal models, so does their sum or difference.** This simple fact is a special property of Normal models and is very important. It allows us to apply our knowledge of Normal probabilities to questions about the sum or difference of independent random variables.



## STEP-BY-STEP EXAMPLE

## Packaging Stereos



Consider the company that manufactures and ships small stereo systems that we just discussed.

Recall that times required to pack the stereos can be described by a Normal model with a mean of 9 minutes and standard deviation of 1.5 minutes. The times for the boxing stage can also be modeled as Normal, with a mean of 6 minutes and standard deviation of 1 minute.

Questions:

1. What is the probability that packing two consecutive systems takes over 20 minutes?
2. What percentage of the stereo systems take longer to pack than to box?

**Question 1:** What is the probability that packing two consecutive systems takes over 20 minutes?



**Plan** State the problem.

**Variables** Define your random variables.

Write an appropriate equation.

Think about the assumptions. Sums of independent Normal random variables follow a Normal model. Such simplicity isn't true in general.

I want to estimate the probability that packing two consecutive systems takes over 20 minutes.

Let  $P_1$  = time for packing the first system

$P_2$  = time for packing the second

$T$  = total time to pack two systems

$$T = P_1 + P_2$$

✓ **Normal Model Assumption:** We are told that both random variables follow Normal models.

✓ **Independence Assumption:** We can reasonably assume that the two packing times are independent.



**Mechanics** Find the expected value.

For sums of independent random variables, variances add. (We don't need the variables to be Normal for this to be true—just independent.)

Find the standard deviation.

Now we use the fact that both random variables follow Normal models to say that their sum is also Normal.

$$\begin{aligned} E(T) &= E(P_1 + P_2) \\ &= E(P_1) + E(P_2) \\ &= 9 + 9 = 18 \text{ minutes} \end{aligned}$$

Since the times are independent,

$$\begin{aligned} \text{Var}(T) &= \text{Var}(P_1 + P_2) \\ &= \text{Var}(P_1) + \text{Var}(P_2) \\ &= 1.5^2 + 1.5^2 \end{aligned}$$

$$\text{Var}(T) = 4.50$$

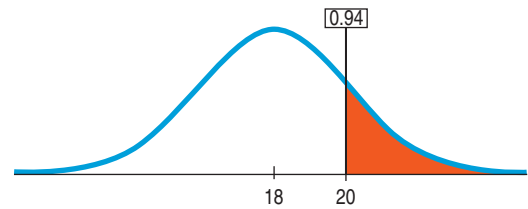
$$\text{SD}(T) = \sqrt{4.50} \approx 2.12 \text{ minutes}$$

I'll model  $T$  with  $N(18, 2.12)$ .

Sketch a picture of the Normal model for the total time, shading the region representing over 20 minutes.

Find the z-score for 20 minutes.

Use technology or Table Z to find the probability.



$$z = \frac{20 - 18}{2.12} = 0.94$$

$$P(T > 20) = P(z > 0.94) = 0.1736$$



**Conclusion** Interpret your result in context.

There's a little more than a 17% chance that it will take a total of over 20 minutes to pack two consecutive stereo systems.

**Question 2: What percentage of the stereo systems take longer to pack than to box?**



**Plan** State the question.

I want to estimate the percentage of the stereo systems that take longer to pack than to box.

**Variables** Define your random variables.

Let  $P$  = time for packing a system  
 $B$  = time for boxing a system  
 $D$  = difference in times to pack and box a system

Write an appropriate equation.

$$D = P - B$$

What are we trying to find? Notice that we can tell which of two quantities is greater by subtracting and asking whether the difference is positive or negative.

The probability that it takes longer to pack than to box a system is the probability that the difference  $P - B$  is greater than zero.

Don't forget to think about the assumptions.

- ✓ **Normal Model Assumption:** We are told that both random variables follow Normal models.
- ✓ **Independence Assumption:** We can assume that the times it takes to pack and to box a system are independent.



**Mechanics** Find the expected value.

$$\begin{aligned} E(D) &= E(P - B) \\ &= E(P) - E(B) \\ &= 9 - 6 = 3 \text{ minutes} \end{aligned}$$

For the difference of independent random variables, variances add.

Find the standard deviation.

State what model you will use.

Sketch a picture of the Normal model for the difference in times, and shade the region representing a difference greater than zero.

Find the z-score for 0 minutes, then use Table Z or technology to find the probability.

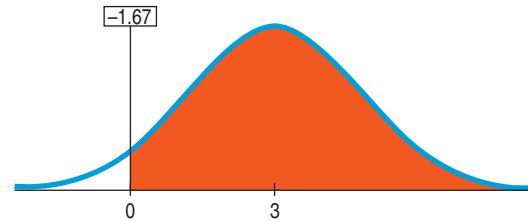
Since the times are independent,

$$\begin{aligned}\text{Var}(D) &= \text{Var}(P - B) \\ &= \text{Var}(P) + \text{Var}(B) \\ &= 1.5^2 + 1^2\end{aligned}$$

$$\text{Var}(D) = 3.25$$

$$\text{SD}(D) = \sqrt{3.25} \approx 1.80 \text{ minutes}$$

I'll model  $D$  with  $N(3, 1.80)$



$$\begin{aligned}z &= \frac{0 - 3}{1.80} = -1.67 \\ P(D > 0) &= P(z > -1.67) = 0.9525\end{aligned}$$



**Conclusion** Interpret your result in context.

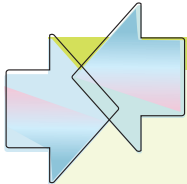
About 95% of all the stereo systems will require more time for packing than for boxing.

## WHAT CAN GO WRONG?

- ▶ **Probability models are still just models.** Models can be useful, but they are not reality. Think about the assumptions behind your models. Are your dice really perfectly fair? (They are probably pretty close.) But when you hear that the probability of a nuclear accident is  $1/10,000,000$  per year, is that likely to be a precise value? Question probabilities as you would data.
- ▶ **If the model is wrong, so is everything else.** Before you try to find the mean or standard deviation of a random variable, check to make sure the probability model is reasonable. As a start, the probabilities in your model should add up to 1. If not, you may have calculated a probability incorrectly or left out a value of the random variable. For instance, in the insurance example, the description mentions only death and disability. Good health is by far the most likely outcome, not to mention the best for both you and the insurance company (who gets to keep your money). Don't overlook that.
- ▶ **Don't assume everything's Normal.** Just because a random variable is continuous or you happen to know a mean and standard deviation doesn't mean that a Normal model will be useful. You must *Think* about whether the **Normality Assumption** is justified. Using a Normal model when it really does not apply will lead to wrong answers and misleading conclusions.

To find the expected value of the sum or difference of random variables, we simply add or subtract means. Center is easy; spread is trickier. Watch out for some common traps.

- ▶ **Watch out for variables that aren't independent.** You can add expected values of *any* two random variables, but you can only add variances of independent random variables. Suppose a survey includes questions about the number of hours of sleep people get each night and also the number of hours they are awake each day. From their answers, we find the mean and standard deviation of hours asleep and hours awake. The expected total must be 24 hours; after all, people are either asleep or awake.<sup>2</sup> The means still add just fine. Since all the totals are exactly 24 hours, however, the standard deviation of the total will be 0. We can't add variances here because the number of hours you're awake depends on the number of hours you're asleep. Be sure to check for independence before adding variances.
- ▶ **Don't forget: Variances of independent random variables add. Standard deviations don't.**
- ▶ **Don't forget: Variances of independent random variables add, even when you're looking at the difference between them.**
- ▶ **Don't write independent instances of a random variable with notation that looks like they are the same variables.** Make sure you write each instance as a different random variable. Just because each random variable describes a similar situation doesn't mean that each random outcome will be the same. These are *random* variables, not the variables you saw in Algebra. Write  $X_1 + X_2 + X_3$  rather than  $X + X + X$ .



## CONNECTIONS

We've seen means, variances, and standard deviations of data. We know that they estimate parameters of models for these data. Now we're looking at the probability models directly. We have only parameters because there are no data to summarize.

It should be no surprise that expected values and standard deviations adjust to shifts and changes of units in the same way as the corresponding data summaries. The fact that we can add variances of independent random quantities is fundamental and will explain why a number of statistical methods work the way they do.

## WHAT HAVE WE LEARNED?



We've learned to work with random variables. We can use the probability model for a discrete random variable to find its expected value and its standard deviation.

We've learned that the mean of the sum or difference of two random variables, discrete or continuous, is just the sum or difference of their means. And we've learned the Pythagorean Theorem of Statistics: *For independent random variables*, the variance of their sum or difference is always the *sum* of their variances.

Finally, we've learned that Normal models are once again special. Sums or differences of Normally distributed random variables also follow Normal models.

<sup>2</sup> Although some students do manage to attain a state of consciousness somewhere between sleeping and wakefulness during Statistics class.

## Terms

Random variable	366. A random variable assumes any of several different numeric values as a result of some random event. Random variables are denoted by a capital letter such as $X$ .
Discrete random variable	366. A random variable that can take one of a finite number <sup>3</sup> of distinct outcomes is called a discrete random variable.
Continuous random variable	366, 367. A random variable that can take any numeric value within a range of values is called a continuous random variable. The range may be infinite or bounded at either or both ends.
Probability model	366. The probability model is a function that associates a probability $P$ with each value of a discrete random variable $X$ , denoted $P(X = x)$ , or with any interval of values of a continuous random variable.
Expected value	367. The expected value of a random variable is its theoretical long-run average value, the center of its model. Denoted $\mu$ or $E(X)$ , it is found (if the random variable is discrete) by summing the products of variable values and probabilities: $\mu = E(X) = \sum xP(x).$
Variance	369. The variance of a random variable is the expected value of the squared deviation from the mean. For discrete random variables, it can be calculated as: $\sigma^2 = \text{Var}(X) = \sum (x - \mu)^2 P(x).$
Standard deviation	369. The standard deviation of a random variable describes the spread in the model, and is the square root of the variance: $\sigma = \text{SD}(X) = \sqrt{\text{Var}(X)}.$
Changing a random variable by a constant:	372. $E(X \pm c) = E(X) \pm c$ $\text{Var}(X \pm c) = \text{Var}(X)$ 373. $E(aX) = aE(X)$ $\text{Var}(aX) = a^2\text{Var}(X)$
Adding or subtracting random variables:	373. $E(X \pm Y) = E(X) \pm E(Y)$ 374. If $X$ and $Y$ are independent, $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y)$ . 374. (The Pythagorean Theorem of Statistics)

## Skills

THINK

- ▶ Be able to recognize random variables.
- ▶ Understand that random variables must be independent in order to determine the variability of their sum or difference by adding variances.

SHOW

- ▶ Be able to find the probability model for a discrete random variable.
- ▶ Know how to find the mean (expected value) and the variance of a random variable.
- ▶ Always use the proper notation for these population parameters:  $\mu$  or  $E(X)$  for the mean, and  $\sigma$ ,  $\text{SD}(X)$ ,  $\sigma^2$ , or  $\text{Var}(X)$  when discussing variability.
- ▶ Know how to determine the new mean and standard deviation after adding a constant, multiplying by a constant, or adding or subtracting two independent random variables.

TELL

- ▶ Be able to interpret the meaning of the expected value and standard deviation of a random variable in the proper context.

<sup>3</sup> Technically, there could be an infinite number of outcomes, as long as they're *countable*. Essentially that means we can imagine listing them all in order, like the counting numbers 1, 2, 3, 4, 5, . . .

## RANDOM VARIABLES ON THE COMPUTER

Statistics packages deal with data, not with random variables. Nevertheless, the calculations needed to find means and standard deviations of random variables are little more than weighted means. Most packages can manage that, but then they are just being overblown calculators. For technological assistance with these calculations, we recommend you pull out your calculator.

## EXERCISES

1. **Expected value.** Find the expected value of each random variable:

a) 

$x$	10	20	30
$P(X=x)$	0.3	0.5	0.2

b) 

$x$	2	4	6	8
$P(X=x)$	0.3	0.4	0.2	0.1

2. **Expected value.** Find the expected value of each random variable:

a) 

$x$	0	1	2
$P(X=x)$	0.2	0.4	0.4

b) 

$x$	100	200	300	400
$P(X=x)$	0.1	0.2	0.5	0.2

3. **Pick a card, any card.** You draw a card from a deck. If you get a red card, you win nothing. If you get a spade, you win \$5. For any club, you win \$10 plus an extra \$20 for the ace of clubs.
- Create a probability model for the amount you win.
  - Find the expected amount you'll win.
  - What would you be willing to pay to play this game?
4. **You bet!** You roll a die. If it comes up a 6, you win \$100. If not, you get to roll again. If you get a 6 the second time, you win \$50. If not, you lose.
- Create a probability model for the amount you win.
  - Find the expected amount you'll win.
  - What would you be willing to pay to play this game?
5. **Kids.** A couple plans to have children until they get a girl, but they agree that they will not have more than three children even if all are boys. (Assume boys and girls are equally likely.)
- Create a probability model for the number of children they might have.
  - Find the expected number of children.
  - Find the expected number of boys they'll have.
6. **Carnival.** A carnival game offers a \$100 cash prize for anyone who can break a balloon by throwing a dart at it. It costs \$5 to play, and you're willing to spend up to \$20 trying to win. You estimate that you have about a 10% chance of hitting the balloon on any throw.

- Create a probability model for this carnival game.
- Find the expected number of darts you'll throw.
- Find your expected winnings.

7. **Software.** A small software company bids on two contracts. It anticipates a profit of \$50,000 if it gets the larger contract and a profit of \$20,000 on the smaller contract. The company estimates there's a 30% chance it will get the larger contract and a 60% chance it will get the smaller contract. Assuming the contracts will be awarded independently, what's the expected profit?

8. **Racehorse.** A man buys a racehorse for \$20,000 and enters it in two races. He plans to sell the horse afterward, hoping to make a profit. If the horse wins both races, its value will jump to \$100,000. If it wins one of the races, it will be worth \$50,000. If it loses both races, it will be worth only \$10,000. The man believes there's a 20% chance that the horse will win the first race and a 30% chance it will win the second one. Assuming that the two races are independent events, find the man's expected profit.

9. **Variation 1.** Find the standard deviations of the random variables in Exercise 1.

10. **Variation 2.** Find the standard deviations of the random variables in Exercise 2.

11. **Pick another card.** Find the standard deviation of the amount you might win drawing a card in Exercise 3.

12. **The die.** Find the standard deviation of the amount you might win rolling a die in Exercise 4.

13. **Kids again.** Find the standard deviation of the number of children the couple in Exercise 5 may have.

14. **Darts.** Find the standard deviation of your winnings throwing darts in Exercise 6.

15. **Repairs.** The probability model below describes the number of repair calls that an appliance repair shop may receive during an hour.

Repair Calls	0	1	2	3
Probability	0.1	0.3	0.4	0.2

- How many calls should the shop expect per hour?
- What is the standard deviation?

16. **Red lights.** A commuter must pass through five traffic lights on her way to work and will have to stop at each one that is red. She estimates the probability model for the number of red lights she hits, as shown below.

$X = \# \text{ of red}$	0	1	2	3	4	5
$P(X = x)$	0.05	0.25	0.35	0.15	0.15	0.05

- a) How many red lights should she expect to hit each day?  
 b) What's the standard deviation?
17. **Defects.** A consumer organization inspecting new cars found that many had appearance defects (dents, scratches, paint chips, etc.). While none had more than three of these defects, 7% had three, 11% two, and 21% one defect. Find the expected number of appearance defects in a new car and the standard deviation.
18. **Insurance.** An insurance policy costs \$100 and will pay policyholders \$10,000 if they suffer a major injury (resulting in hospitalization) or \$3000 if they suffer a minor injury (resulting in lost time from work). The company estimates that each year 1 in every 2000 policyholders may have a major injury, and 1 in 500 a minor injury only.
- a) Create a probability model for the profit on a policy.  
 b) What's the company's expected profit on this policy?  
 c) What's the standard deviation?
19. **Cancelled flights.** Mary is deciding whether to book the cheaper flight home from college after her final exams, but she's unsure when her last exam will be. She thinks there is only a 20% chance that the exam will be scheduled after the last day she can get a seat on the cheaper flight. If it is and she has to cancel the flight, she will lose \$150. If she can take the cheaper flight, she will save \$100.
- a) If she books the cheaper flight, what can she expect to gain, on average?  
 b) What is the standard deviation?
20. **Day trading.** An option to buy a stock is priced at \$200. If the stock closes above 30 on May 15, the option will be worth \$1000. If it closes below 20, the option will be worth nothing, and if it closes between 20 and 30 (inclusively), the option will be worth \$200. A trader thinks there is a 50% chance that the stock will close in the 20–30 range, a 20% chance that it will close above 30, and a 30% chance that it will fall below 20 on May 15.
- a) Should she buy the stock option?  
 b) How much does she expect to gain?  
 c) What is the standard deviation of her gain?
21. **Contest.** You play two games against the same opponent. The probability you win the first game is 0.4. If you win the first game, the probability you also win the second is 0.2. If you lose the first game, the probability that you win the second is 0.3.
- a) Are the two games independent? Explain.  
 b) What's the probability you lose both games?  
 c) What's the probability you win both games?  
 d) Let random variable  $X$  be the number of games you win. Find the probability model for  $X$ .  
 e) What are the expected value and standard deviation?

22. **Contracts.** Your company bids for two contracts. You believe the probability you get contract #1 is 0.8. If you get contract #1, the probability you also get contract #2 will be 0.2, and if you do not get #1, the probability you get #2 will be 0.3.
- a) Are the two contracts independent? Explain.  
 b) Find the probability you get both contracts.  
 c) Find the probability you get no contract.  
 d) Let  $X$  be the number of contracts you get. Find the probability model for  $X$ .  
 e) Find the expected value and standard deviation.
23. **Batteries.** In a group of 10 batteries, 3 are dead. You choose 2 batteries at random.
- a) Create a probability model for the number of good batteries you get.  
 b) What's the expected number of good ones you get?  
 c) What's the standard deviation?
24. **Kittens.** In a litter of seven kittens, three are female. You pick two kittens at random.
- a) Create a probability model for the number of male kittens you get.  
 b) What's the expected number of males?  
 c) What's the standard deviation?

25. **Random variables.** Given independent random variables with means and standard deviations as shown, find the mean and standard deviation of:

	Mean	SD
$X$	10	2
$Y$	20	5

- a)  $3X$   
 b)  $Y + 6$   
 c)  $X + Y$   
 d)  $X - Y$   
 e)  $X_1 + X_2$

26. **Random variables.** Given independent random variables with means and standard deviations as shown, find the mean and standard deviation of:

	Mean	SD
$X$	80	12
$Y$	12	3

- a)  $X - 20$   
 b)  $0.5Y$   
 c)  $X + Y$   
 d)  $X - Y$   
 e)  $Y_1 + Y_2$

27. **Random variables.** Given independent random variables with means and standard deviations as shown, find the mean and standard deviation of:

	Mean	SD
$X$	120	12
$Y$	300	16

- a)  $0.8Y$   
 b)  $2X - 100$   
 c)  $X + 2Y$   
 d)  $3X - Y$   
 e)  $Y_1 + Y_2$

28. **Random variables.** Given independent random variables with means and standard deviations as shown, find the mean and standard deviation of:

	Mean	SD
$X$	80	12
$Y$	12	3

- a)  $2Y + 20$   
 b)  $3X$   
 c)  $0.25X + Y$   
 d)  $X - 5Y$   
 e)  $X_1 + X_2 + X_3$

29. **Eggs.** A grocery supplier believes that in a dozen eggs, the mean number of broken ones is 0.6 with a standard

- deviation of 0.5 eggs. You buy 3 dozen eggs without checking them.
- How many broken eggs do you expect to get?
  - What's the standard deviation?
  - What assumptions did you have to make about the eggs in order to answer this question?
30. **Garden.** A company selling vegetable seeds in packets of 20 estimates that the mean number of seeds that will actually grow is 18, with a standard deviation of 1.2 seeds. You buy 5 different seed packets.
- How many bad seeds do you expect to get?
  - What's the standard deviation?
  - What assumptions did you make about the seeds? Do you think that assumption is warranted? Explain.
31. **Repair calls.** Find the mean and standard deviation of the number of repair calls the appliance shop in Exercise 15 should expect during an 8-hour day.
32. **Stop!** Find the mean and standard deviation of the number of red lights the commuter in Exercise 16 should expect to hit on her way to work during a 5-day work week.
33. **Tickets.** A delivery company's trucks occasionally get parking tickets, and based on past experience, the company plans that the trucks will average 1.3 tickets a month, with a standard deviation of 0.7 tickets.
- If they have 18 trucks, what are the mean and standard deviation of the total number of parking tickets the company will have to pay this month?
  - What assumption did you make in answering?
34. **Donations.** Organizers of a televised fundraiser know from past experience that most people donate small amounts (\$10–\$25), some donate larger amounts (\$50–\$100), and a few people make very generous donations of \$250, \$500, or more. Historically, pledges average about \$32 with a standard deviation of \$54.
- If 120 people call in pledges, what are the mean and standard deviation of the total amount raised?
  - What assumption did you make in answering this question?
35. **Fire!** An insurance company estimates that it should make an annual profit of \$150 on each homeowner's policy written, with a standard deviation of \$6000.
- Why is the standard deviation so large?
  - If it writes only two of these policies, what are the mean and standard deviation of the annual profit?
  - If it writes 10,000 of these policies, what are the mean and standard deviation of the annual profit?
  - Is the company likely to be profitable? Explain.
  - What assumptions underlie your analysis? Can you think of circumstances under which those assumptions might be violated? Explain.
36. **Casino.** A casino knows that people play the slot machines in hopes of hitting the jackpot but that most of them lose their dollar. Suppose a certain machine pays out an average of \$0.92, with a standard deviation of \$120.
- Why is the standard deviation so large?
  - If you play 5 times, what are the mean and standard deviation of the casino's profit?
  - If gamblers play this machine 1000 times in a day, what are the mean and standard deviation of the casino's profit?
  - Is the casino likely to be profitable? Explain.
37. **Cereal.** The amount of cereal that can be poured into a small bowl varies with a mean of 1.5 ounces and a standard deviation of 0.3 ounces. A large bowl holds a mean of 2.5 ounces with a standard deviation of 0.4 ounces. You open a new box of cereal and pour one large and one small bowl.
- How much more cereal do you expect to be in the large bowl?
  - What's the standard deviation of this difference?
  - If the difference follows a Normal model, what's the probability the small bowl contains more cereal than the large one?
  - What are the mean and standard deviation of the total amount of cereal in the two bowls?
  - If the total follows a Normal model, what's the probability you poured out more than 4.5 ounces of cereal in the two bowls together?
  - The amount of cereal the manufacturer puts in the boxes is a random variable with a mean of 16.3 ounces and a standard deviation of 0.2 ounces. Find the expected amount of cereal left in the box and the standard deviation.
38. **Pets.** The American Veterinary Association claims that the annual cost of medical care for dogs averages \$100, with a standard deviation of \$30, and for cats averages \$120, with a standard deviation of \$35.
- What's the expected difference in the cost of medical care for dogs and cats?
  - What's the standard deviation of that difference?
  - If the costs can be described by Normal models, what's the probability that medical expenses are higher for someone's dog than for her cat?
  - What concerns do you have?
39. **More cereal.** In Exercise 37 we poured a large and a small bowl of cereal from a box. Suppose the amount of cereal that the manufacturer puts in the boxes is a random variable with mean 16.2 ounces and standard deviation 0.1 ounces.
- Find the expected amount of cereal left in the box.
  - What's the standard deviation?
  - If the weight of the remaining cereal can be described by a Normal model, what's the probability that the box still contains more than 13 ounces?
40. **More pets.** You're thinking about getting two dogs and a cat. Assume that annual veterinary expenses are independent and have a Normal model with the means and standard deviations described in Exercise 38.
- Define appropriate variables and express the total annual veterinary costs you may have.
  - Describe the model for this total cost. Be sure to specify its name, expected value, and standard deviation.
  - What's the probability that your total expenses will exceed \$400?
41. **Medley.** In the  $4 \times 100$  medley relay event, four swimmers swim 100 yards, each using a different stroke. A



college team preparing for the conference championship looks at the times their swimmers have posted and creates a model based on the following assumptions:

- The swimmers' performances are independent.
- Each swimmer's times follow a Normal model.
- The means and standard deviations of the times (in seconds) are as shown:

Swimmer	Mean	SD
1 (backstroke)	50.72	0.24
2 (breaststroke)	55.51	0.22
3 (butterfly)	49.43	0.25
4 (freestyle)	44.91	0.21

- a) What are the mean and standard deviation for the relay team's total time in this event?
- b) The team's best time so far this season was 3:19.48. (That's 199.48 seconds.) Do you think the team is likely to swim faster than this at the conference championship? Explain.
42. **Bikes.** Bicycles arrive at a bike shop in boxes. Before they can be sold, they must be unpacked, assembled, and tuned (lubricated, adjusted, etc.). Based on past experience, the shop manager makes the following assumptions about how long this may take:
- The times for each setup phase are independent.
  - The times for each phase follow a Normal model.
  - The means and standard deviations of the times (in minutes) are as shown:

Phase	Mean	SD
Unpacking	3.5	0.7
Assembly	21.8	2.4
Tuning	12.3	2.7

- a) What are the mean and standard deviation for the total bicycle setup time?
- b) A customer decides to buy a bike like one of the display models but wants a different color. The shop has one, still in the box. The manager says they can have it ready in half an hour. Do you think the bike will be set up and ready to go as promised? Explain.
43. **Farmers' market.** A farmer has 100 lb of apples and 50 lb of potatoes for sale. The market price for apples (per pound) each day is a random variable with a mean of 0.5 dollars and a standard deviation of 0.2 dollars. Similarly, for a pound of potatoes, the mean price is 0.3 dollars and the standard deviation is 0.1 dollars. It also costs him 2 dollars to bring all the apples and potatoes to the market. The market is busy with eager shoppers, so we can assume that he'll be able to sell all of each type of produce at that day's price.
- a) Define your random variables, and use them to express the farmer's net income.
- b) Find the mean.
- c) Find the standard deviation of the net income.
- d) Do you need to make any assumptions in calculating the mean? How about the standard deviation?

44. **Bike sale.** The bicycle shop in Exercise 42 will be offering 2 specially priced children's models at a sidewalk sale. The basic model will sell for \$120 and the deluxe model for \$150. Past experience indicates that sales of the basic model will have a mean of 5.4 bikes with a standard deviation of 1.2, and sales of the deluxe model will have a mean of 3.2 bikes with a standard deviation of 0.8 bikes. The cost of setting up for the sidewalk sale is \$200.
- a) Define random variables and use them to express the bicycle shop's net income.
- b) What's the mean of the net income?
- c) What's the standard deviation of the net income?
- d) Do you need to make any assumptions in calculating the mean? How about the standard deviation?
45. **Coffee and doughnuts.** At a certain coffee shop, all the customers buy a cup of coffee; some also buy a doughnut. The shop owner believes that the number of cups he sells each day is normally distributed with a mean of 320 cups and a standard deviation of 20 cups. He also believes that the number of doughnuts he sells each day is independent of the coffee sales and is normally distributed with a mean of 150 doughnuts and a standard deviation of 12.
- a) The shop is open every day but Sunday. Assuming day-to-day sales are independent, what's the probability he'll sell over 2000 cups of coffee in a week?
- b) If he makes a profit of 50 cents on each cup of coffee and 40 cents on each doughnut, can he reasonably expect to have a day's profit of over \$300? Explain.
- c) What's the probability that on any given day he'll sell a doughnut to more than half of his coffee customers?
46. **Weightlifting.** The Atlas BodyBuilding Company (ABC) sells "starter sets" of barbells that consist of one bar, two 20-pound weights, and four 5-pound weights. The bars weigh an average of 10 pounds with a standard deviation of 0.25 pounds. The weights average the specified amounts, but the standard deviations are 0.2 pounds for the 20-pounders and 0.1 pounds for the 5-pounders. We can assume that all the weights are normally distributed.
- a) ABC ships these starter sets to customers in two boxes: The bar goes in one box and the six weights go in another. What's the probability that the total weight in that second box exceeds 60.5 pounds? Define your variables clearly and state any assumptions you make.
- b) It costs ABC \$0.40 per pound to ship the box containing the weights. Because it's an odd-shaped package, though, shipping the bar costs \$0.50 a pound plus a \$6.00 surcharge. Find the mean and standard deviation of the company's total cost for shipping a starter set.
- c) Suppose a customer puts a 20-pound weight at one end of the bar and the four 5-pound weights at the other end. Although he expects the two ends to weigh the same, they might differ slightly. What's the probability the difference is more than a quarter of a pound?



## **JUST CHECKING**

### **Answers**

1. a)

Outcome	$X = \text{cost}$	Probability
Recharging works	\$60	0.75
Replace control unit	\$200	0.25

b)  $60(0.75) + 200(0.25) = \$95$

c) Car owners with this problem will spend an average of \$95 to get it fixed.

2. a)  $100 + 100 = 200$  seconds

b)  $\sqrt{50^2 + 50^2} = 70.7$  seconds

c) The times for the two customers are independent.

# Probability Models



Suppose a cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal, in the hope of increasing sales. The manufacturer announces that 20% of the boxes contain a picture of Tiger Woods, 30% a picture of David Beckham, and the rest a picture of Serena Williams.

Sound familiar? In Chapter 11 we simulated to find the number of boxes we'd need to open to get one of each card. That's a fairly complex question and one well suited for simulation. But many important questions can be answered more directly by using simple probability models.

## Searching for Tiger



You're a huge Tiger Woods fan. You don't care about completing the whole sports card collection, but you've just *got* to have the Tiger Woods picture. How many boxes do you expect you'll have to open before you find him? This isn't the same question that we asked before, but this situation is simple enough for a probability model.

We'll keep the assumption that pictures are distributed at random and we'll trust the manufacturer's claim that 20% of the cards are Tiger. So, when you open the box, the probability that you succeed in finding Tiger is 0.20. Now we'll call the act of opening *each* box a trial, and note that:

- ▶ There are only two possible outcomes (called *success* and *failure*) on each trial. Either you get Tiger's picture (success), or you don't (failure).
- ▶ In advance, the probability of success, denoted  $p$ , is the same on every trial. Here  $p = 0.20$  for each box.
- ▶ As we proceed, the trials are independent. Finding Tiger in the first box does not change what might happen when you reach for the next box.

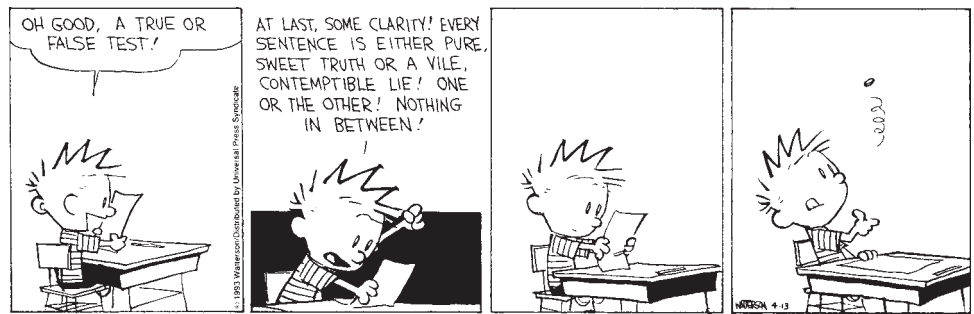
Situations like this occur often, and are called **Bernoulli trials**. Common examples of Bernoulli trials include tossing a coin, looking for defective products rolling off an assembly line, or even shooting free throws in a basketball game. Just as we found equally likely random digits to be the building blocks for our simulation, we can use Bernoulli trials to build a wide variety of useful probability models.



Daniel Bernoulli (1700–1782) was the nephew of Jacob, whom you saw in Chapter 14. He was the first to work out the mathematics for what we now call Bernoulli trials.

### AS Activity: Bernoulli Trials.

Guess what! We've been generating Bernoulli trials all along. Look at the Random Simulation Tool in a new way.



Calvin and Hobbes © 1993 Watterson. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

Back to Tiger. We want to know how many boxes we'll need to open to find his card. Let's call this random variable  $Y = \# \text{ boxes}$ , and build a probability model for it. What's the probability you find his picture in the first box of cereal? It's 20%, of course. We could write  $P(Y = 1) = 0.20$ .

How about the probability that you don't find Tiger until the second box? Well, that means you fail on the first trial and then succeed on the second. With the probability of success 20%, the probability of failure, denoted  $q$ , is  $1 - 0.2 = 80\%$ . Since the trials are independent, the probability of getting your first success on the second trial is  $P(Y = 2) = (0.8)(0.2) = 0.16$ .

Of course, you could have a run of bad luck. Maybe you won't find Tiger until the fifth box of cereal. What are the chances of that? You'd have to fail 4 straight times and then succeed, so  $P(Y = 5) = (0.8)^4(0.2) = 0.08192$ .

How many boxes might you expect to have to open? We could reason that since Tiger's picture is in 20% of the boxes, or 1 in 5, we expect to find his picture, on average, in the fifth box; that is,  $E(Y) = \frac{1}{0.2} = 5$  boxes. That's correct, but not easy to prove.

## The Geometric Model

### TI-*nspire*

**Geometric probabilities.** See what happens to a geometric model as you change the probability of success.

### NOTATION ALERT:

Now we have two more reserved letters. Whenever we deal with Bernoulli trials,  $p$  represents the probability of success, and  $q$  the probability of failure. (Of course,  $q = 1 - p$ .)

We want to model how long it will take to achieve the first success in a series of Bernoulli trials. The model that tells us this probability is called the **Geometric probability model**. Geometric models are completely specified by one parameter,  $p$ , the probability of success, and are denoted  $\text{Geom}(p)$ . Since achieving the first success on trial number  $x$  requires first experiencing  $x - 1$  failures, the probabilities are easily expressed by a formula.

### GEOMETRIC PROBABILITY MODEL FOR BERNOULLI TRIALS: $\text{Geom}(p)$

$p$  = probability of success (and  $q = 1 - p$  = probability of failure)

$X$  = number of trials until the first success occurs

$$P(X = x) = q^{x-1}p$$

Expected value:  $E(X) = \mu = \frac{1}{p}$

Standard deviation:  $\sigma = \sqrt{\frac{q}{p^2}}$

## FOR EXAMPLE

## Spam and the Geometric model

*Postini* is a global company specializing in communications security. The company monitors over 1 billion Internet messages per day and recently reported that 91% of e-mails are spam!

Let's assume that your e-mail is typical—91% spam. We'll also assume you aren't using a spam filter, so every message gets dumped in your inbox. And, since spam comes from many different sources, we'll consider your messages to be independent.

**Questions:** Overnight your inbox collects e-mail. When you first check your e-mail in the morning, about how many spam e-mails should you expect to have to wade through and discard before you find a real message? What's the probability that the 4th message in your inbox is the first one that isn't spam?

There are two outcomes: a real message (success) and spam (failure). Since 91% of e-mails are spam, the probability of success  $p = 1 - 0.91 = 0.09$ .

Let  $X$  = the number of e-mails I'll check until I find a real message. I assume that the messages arrive independently and in a random order. I can use the model  $\text{Geom}(0.09)$ .

$$E(X) = \frac{1}{p} = \frac{1}{0.09} = 11.1$$

$$P(X = 4) = (0.91)^3(0.09) = 0.0678$$

On average, I expect to have to check just over 11 e-mails before I find a real message. There's slightly less than a 7% chance that my first real message will be the 4th one I check.

Note that the probability calculation isn't new. It's simply Chapter 14's Multiplication Rule used to find  $P(\text{spam} \cap \text{spam} \cap \text{spam} \cap \text{real})$ .

## MATH BOX

We want to find the mean (expected value) of random variable  $X$ , using a geometric model with probability of success  $p$ .

First, write the probabilities:

$x$	1	2	3	4	...
$P(X = x)$	$p$	$qp$	$q^2p$	$q^3p$	...

The expected value is:

$$E(X) = 1p + 2qp + 3q^2p + 4q^3p + \dots$$

Let  $p = 1 - q$ :

$$= (1 - q) + 2q(1 - q) + 3q^2(1 - q) + 4q^3(1 - q) + \dots$$

Simplify:

$$= 1 - q + 2q - 2q^2 + 3q^2 - 3q^3 + 4q^3 - 4q^4 + \dots$$

That's an infinite geometric series, with first term 1 and common ratio  $q$ :

$$= 1 + q + q^2 + q^3 + \dots$$

$$= \frac{1}{1 - q}$$

So, finally . . .

$$E(X) = \frac{1}{p}$$

## Independence

One of the important requirements for Bernoulli trials is that the trials be independent. Sometimes that's a reasonable assumption—when tossing a coin or rolling a die, for example. But that becomes a problem when (often!) we're looking at situations involving samples chosen without replacement. We said that whether we find a Tiger Woods card in one box has no effect on the probabilities

in other boxes. This is *almost* true. Technically, if exactly 20% of the boxes have Tiger Woods cards, then when you find one, you've reduced the number of remaining Tiger Woods cards. If you knew there were 2 Tiger Woods cards hiding in the 10 boxes of cereal on the market shelf, then finding one in the first box you try would clearly change your chances of finding Tiger in the next box. With a few million boxes of cereal, though, the difference is hardly worth mentioning.

If we had an infinite number of boxes, there wouldn't be a problem. It's selecting from a finite population that causes the probabilities to change, making the trials not independent. Obviously, taking 2 out of 10 boxes changes the probability. Taking even a few hundred out of millions, though, makes very little difference. Fortunately, we have a rule of thumb for the in-between cases. It turns out that if we look at less than 10% of the population, we can pretend that the trials are independent and still calculate probabilities that are quite accurate.

**The 10% Condition:** Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.

## STEP-BY-STEP EXAMPLE

### Working with a Geometric Model

People with O-negative blood are called “universal donors” because O-negative blood can be given to anyone else, regardless of the recipient's blood type. Only about 6% of people have O-negative blood.

#### Questions:

1. If donors line up at random for a blood drive, how many do you expect to examine before you find someone who has O-negative blood?
2. What's the probability that the first O-negative donor found is one of the first four people in line?



**Plan** State the questions.

Check to see that these are Bernoulli trials.

**Variable** Define the random variable.

**Model** Specify the model.

I want to estimate how many people I'll need to check to find an O-negative donor, and the probability that 1 of the first 4 people is O-negative.

- ✓ There are two outcomes:  
     *success = O-negative*  
     *failure = other blood types*
- ✓ The probability of *success* for each person is  $p = 0.06$ , because they lined up randomly.
- ✓ **10% Condition:** Trials aren't independent because the population is finite, but the donors lined up are fewer than 10% of all possible donors.

Let  $X$  = number of donors until one is O-negative.

I can model  $X$  with  $\text{Geom}(0.06)$ .

**Mechanics** Find the mean.

Calculate the probability of success on one of the first four trials. That's the probability that  $X = 1, 2, 3,$  or  $4$ .

$$E(X) = \frac{1}{0.06} \approx 16.7$$

$$\begin{aligned} P(X \leq 4) &= P(X = 1) + P(X = 2) + \\ &\quad P(X = 3) + P(X = 4) \\ &= (0.06) + (0.94)(0.06) + \\ &\quad (0.94)^2(0.06) + (0.94)^3(0.06) \\ &\approx 0.2193 \end{aligned}$$

**Conclusion** Interpret your results in context.

Blood drives such as this one expect to examine an average of 16.7 people to find a universal donor. About 22% of the time there will be one within the first 4 people in line.

## TI TIPS

## Finding geometric probabilities

```

DISTR DRAW
G:Fcdf(
A:binompdf(
B:binomcdf(
C:Poissonpdf(
D:Poissoncdf(
E:Geometpdf(
F:Geometcdf(

```

```

Geometpdf(.2,5)
.08192

```

```

Geometcdf(.2,4)
.5904

```

Your TI knows the geometric model. Just as you saw back in Chapter 6 with the Normal model, commands to calculate probability distributions are found in the 2nd DISTR menu. Have a look. After many others (Don't drop the course yet!) you'll see two Geometric probability functions at the bottom of the list.

- **Geometpdf(.**

The "pdf" stands for "probability density function." This command allows you to find the probability of any *individual* outcome. You need only specify  $p$ , which defines the Geometric model, and  $x$ , which indicates the number of trials until you get a success. The format is **Geometpdf( $p, x$ )**.

For example, suppose we want to know the probability that we find our first Tiger Woods picture in the fifth box of cereal. Since Tiger is in 20% of the boxes, we use  $p = 0.2$  and  $x = 5$ , entering the command **Geometpdf(.2,5)**. The calculator says there's about an 8% chance.

- **Geometcdf(.**

This is the "cumulative density function," meaning that it finds the sum of the probabilities of several possible outcomes. In general, the command **Geometcdf( $p, x$ )** calculates the probability of finding the first success *on or before* the  $x$ th trial.

Let's find the probability of getting a Tiger Woods picture by the time we open the fourth box of cereal—in other words, the probability our first success comes on the first box, or the second, or the third, or the fourth. Again we specify  $p = 0.2$ , and now use  $x = 4$ . The command **Geometcdf(.2,4)** calculates all the probabilities and adds them. There's about a 59% chance that our quest for a Tiger Woods photo will succeed by the time we open the fourth box.

## The Binomial Model

We can use the Bernoulli trials to answer other questions. Suppose you buy 5 boxes of cereal. What's the probability you get *exactly* 2 pictures of Tiger Woods? Before, we asked how long it would take until our first success. Now we want to find the probability of getting 2 successes among the 5 trials. We are still talking about Bernoulli trials, but we're asking a different question.

**A S** **Activity: The Binomial Distribution.** It's more interesting to combine Bernoulli trials. Simulate this with the Random Tool to get a sense of how Binomial models behave.

This time we're interested in the *number of successes* in the 5 trials, so we'll call it  $X = \text{number of successes}$ . We want to find  $P(X = 2)$ . This is an example of a **Binomial probability**. It takes two parameters to define this **Binomial model**: the number of trials,  $n$ , and the probability of success,  $p$ . We denote this model  $\text{Binom}(n, p)$ . Here,  $n = 5$  trials, and  $p = 0.2$ , the probability of finding a Tiger Woods card in any trial.

Exactly 2 successes in 5 trials means 2 successes and 3 failures. It seems logical that the probability should be  $(0.2)^2(0.8)^3$ . Too bad! It's not that easy. That calculation would give you the probability of finding Tiger in the first 2 boxes and not in the next 3—in that order. But you could find Tiger in the third and fifth boxes and still have 2 successes. The probability of those outcomes in that particular order is  $(0.8)(0.8)(0.2)(0.8)(0.2)$ . That's also  $(0.2)^2(0.8)^3$ . In fact, the probability will always be the same, no matter what order the successes and failures occur in. Anytime we get 2 successes in 5 trials, no matter what the order, the probability will be  $(0.2)^2(0.8)^3$ . We just need to take account of all the possible orders in which the outcomes can occur.

Fortunately, these possible orders are *disjoint*. (For example, if your two successes came on the first two trials, they couldn't come on the last two.) So we can use the Addition Rule and add up the probabilities for all the possible orderings. Since the probabilities are all the same, we only need to know how many orders are possible. For small numbers, we can just make a tree diagram and count the branches. For larger numbers this isn't practical, so we let the computer or calculator do the work.

Each different order in which we can have  $k$  successes in  $n$  trials is called a "combination." The total number of ways that can happen is written  $\binom{n}{k}$  or  ${}_nC_k$  and pronounced "n choose k."

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \text{ where } n! \text{ (pronounced "n factorial")} = n \times (n-1) \times \cdots \times 1$$

For 2 successes in 5 trials,

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = \frac{5 \times 4}{2 \times 1} = 10.$$

So there are 10 ways to get 2 Tiger pictures in 5 boxes, and the probability of each is  $(0.2)^2(0.8)^3$ . Now we can find what we wanted:

$$P(\#\text{success} = 2) = 10(0.2)^2(0.8)^3 = 0.2048$$

In general, the probability of exactly  $k$  successes in  $n$  trials is  $\binom{n}{k} p^k q^{n-k}$ .

Using this formula, we could find the expected value by adding up  $xP(X = x)$  for all values, but it would be a long, hard way to get an answer that you already know intuitively. What's the expected value? If we have 5 boxes, and Tiger's picture is in 20% of them, then we would expect to have  $5(0.2) = 1$  success. If we had 100 trials with probability of success 0.2, how many successes would you expect? Can you think of any reason not to say 20? It seems so simple that most people wouldn't even stop to think about it. You just multiply the probability of success by  $n$ . In other words,  $E(X) = np$ . Not fully convinced? We prove it in the next Math Box.

The standard deviation is less obvious; you can't just rely on your intuition. Fortunately, the formula for the standard deviation also boils down to something simple:  $SD(X) = \sqrt{npq}$ . (If you're curious about where that comes from, it's in the Math Box too!) In 100 boxes of cereal, we expect to find 20 Tiger Woods cards, with a standard deviation of  $\sqrt{100 \times 0.8 \times 0.2} = 4$  pictures.

Time to summarize. A **Binomial probability model** describes the number of successes in a specified number of trials. It takes two parameters to specify this model: the number of trials  $n$  and the probability of success  $p$ .

#### NOTATION ALERT:

Now punctuation! Throughout mathematics  $n!$ , pronounced "n factorial," is the product of all the integers from 1 to  $n$ . For example,  $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$ .



TI-*n*spire**Binomial probabilities.**

Do-it-yourself binomial models!  
Watch the probabilities change as  
you control  $n$  and  $p$ .

**BINOMIAL PROBABILITY MODEL FOR BERNOULLI TRIALS: Binom( $n, p$ )**

$n$  = number of trials

$p$  = probability of success (and  $q = 1 - p$  = probability of failure)

$X$  = number of successes in  $n$  trials

$$P(X = x) = {}_n C_x p^x q^{n-x}, \text{ where } {}_n C_x = \frac{n!}{x!(n-x)!}$$

Mean:  $\mu = np$

Standard Deviation:  $\sigma = \sqrt{npq}$

**MATH BOX**

To derive the formulas for the mean and standard deviation of a Binomial model we start with the most basic situation.

Consider a single Bernoulli trial with probability of success  $p$ . Let's find the mean and variance of the number of successes.

Here's the probability model  
for the number of successes:

$x$	0	1
$P(X = x)$	$q$	$p$

Find the expected value:

$$E(X) = 0q + 1p$$

$$E(X) = p$$

And now the variance:

$$\text{Var}(X) = (0 - p)^2 q + (1 - p)^2 p$$

$$= p^2 q + q^2 p$$

$$= pq(p + q)$$

$$= pq(1)$$

$$\text{Var}(X) = pq$$

What happens when there is more than one trial, though? A Binomial model simply counts the number of successes in a series of  $n$  independent Bernoulli trials. That makes it easy to find the mean and standard deviation of a binomial random variable,  $Y$ .

$$\text{Let } Y = X_1 + X_2 + X_3 + \cdots + X_n$$

$$E(Y) = E(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$= E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_n)$$

$$= p + p + p + \cdots + p \text{ (There are } n \text{ terms.)}$$

So, as we thought, the mean is  $E(Y) = np$ .

And since the trials are independent, the variances add:

$$\text{Var}(Y) = \text{Var}(X_1 + X_2 + X_3 + \cdots + X_n)$$

$$= \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + \cdots + \text{Var}(X_n)$$

$$= pq + pq + pq + \cdots + pq \text{ (Again, } n \text{ terms.)}$$

$$\text{Var}(Y) = npq$$

Voilà! The standard deviation is  $SD(Y) = \sqrt{npq}$ .

## FOR EXAMPLE

## Spam and the Binomial model

**Recap:** The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Suppose your inbox contains 25 messages.

**Questions:** What are the mean and standard deviation of the number of real messages you should expect to find in your inbox? What's the probability that you'll find only 1 or 2 real messages?

I assume that messages arrive independently and at random, with the probability of success (a real message)  $p = 1 - 0.91 = 0.09$ . Let  $X$  = the number of real messages among 25. I can use the model  $\text{Binom}(25, 0.09)$ .

$$\begin{aligned} E(X) &= np = 25(0.09) = 2.25 \\ SD(X) &= \sqrt{npq} = \sqrt{25(0.09)(0.91)} = 1.43 \\ P(X = 1 \text{ or } 2) &= P(X = 1) + P(X = 2) \\ &= \binom{25}{1}(0.09)^1(0.91)^{24} + \binom{25}{2}(0.09)^2(0.91)^{23} \\ &= 0.2340 + 0.2777 \\ &= 0.5117 \end{aligned}$$

Among 25 e-mail messages, I expect to find an average of 2.25 that aren't spam, with a standard deviation of 1.43 messages. There's just over a 50% chance that 1 or 2 of my 25 e-mails will be real messages.

## STEP-BY-STEP EXAMPLE

## Working with a Binomial Model

Suppose 20 donors come to a blood drive. Recall that 6% of people are "universal donors."

**Questions:**

1. What are the mean and standard deviation of the number of universal donors among them?
2. What is the probability that there are 2 or 3 universal donors?



**Plan** State the question.

Check to see that these are Bernoulli trials.

**Variable** Define the random variable.

**Model** Specify the model.

I want to know the mean and standard deviation of the number of universal donors among 20 people, and the probability that there are 2 or 3 of them.

✓ There are two outcomes:



success = O-negative  
failure = other blood types

✓  $p = 0.06$ , because people have lined up at random.

✓ **10% Condition:** Trials are not independent, because the population is finite, but fewer than 10% of all possible donors are lined up.

Let  $X$  = number of O-negative donors among  $n = 20$  people.

I can model  $X$  with  $\text{Binom}(20, 0.06)$ .

 <p><b>Mechanics</b> Find the expected value and standard deviation.</p>	$E(X) = np = 20(0.06) = 1.2$ $SD(X) = \sqrt{npq} = \sqrt{20(0.06)(0.94)} \approx 1.06$ $P(X = 2 \text{ or } 3) = P(X = 2) + P(X = 3)$ $= \binom{20}{2}(0.06)^2(0.94)^{18}$ $+ \binom{20}{3}(0.06)^3(0.94)^{17}$ $\approx 0.2246 + 0.0860$ $= 0.3106$
 <p><b>Conclusion</b> Interpret your results in context.</p>	<p>In groups of 20 randomly selected blood donors, I expect to find an average of 1.2 universal donors, with a standard deviation of 1.06. About 31% of the time, I'd find 2 or 3 universal donors among the 20 people.</p>

### TI Tips

### Finding binomial probabilities

```
DISTR DRAW
D:dfcdf(
B:binompdf(
B:binomcdf(
C:Poissonpdf(
D:Poissoncdf(
E:geometpdf(
F:geometcdf(
```

```
binompdf(5,.2,2)
.2048
```

```
binomcdf(10,.2,4)
.9672065025
```

```
1-binomcdf(10,.2,3)
.1208738816
```

Remember how the calculator handles Geometric probabilities? Well, the commands for finding Binomial probabilities are essentially the same. Again you'll find them in the 2nd DISTR menu.

- **binompdf(**

This probability density function allows you to find the probability of an *individual* outcome. You need to define the Binomial model by specifying  $n$  and  $p$ , and then indicate the desired number of successes,  $x$ . The format is **binompdf( $n, p, X$ )**.

For example, recall that Tiger Woods' picture is in 20% of the cereal boxes. Suppose that we want to know the probability of finding Tiger exactly twice among 5 boxes of cereal. We use  $n = 5, p = 0.2$ , and  $x = 2$ , entering the command **binompdf(5, .2, 2)**. There's about a 20% chance of getting two pictures of Tiger Woods in five boxes of cereal.

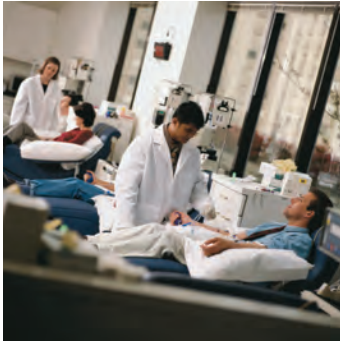
- **binomcdf(**

Need to add several Binomial probabilities? To find the total probability of getting  $x$  or fewer successes among the  $n$  trials use the cumulative Binomial density function **binomcdf( $n, p, X$ )**.

For example, suppose we have ten boxes of cereal and wonder about the probability of finding up to 4 pictures of Tiger. That's the probability of 0, 1, 2, 3 or 4 successes, so we specify the command **binomcdf(10, .2, 4)**. Pretty likely!

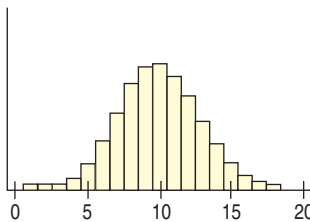
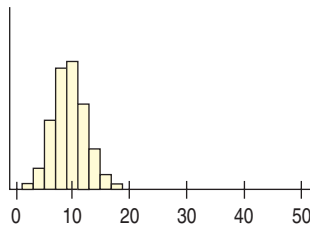
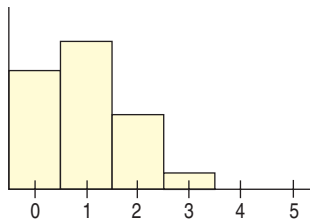
Of course "up to 4" allows for the possibility that we end up with none. What's the probability we get at least 4 pictures of Tiger in 10 boxes? Well, "at least 4" means "not 3 or fewer." That's the complement of 0, 1, 2, or 3 successes. Have your TI evaluate **1-binomcdf(10, .2, 3)**. There's about a 12% chance we'll find at least 4 pictures of Tiger in 10 boxes of cereal.

## The Normal Model to the Rescue!



**A S**

**Activity: Normal Approximation.** Binomial probabilities can be hard to calculate. With the Simulation Tool you'll see how well the Normal model can approximate the Binomial—a much easier method.



### TI-*n*spire

**How close to Normal?** How well does a Normal curve fit a binomial model? Check out the Success/Failure Condition for yourself.

Suppose the Tennessee Red Cross anticipates the need for at least 1850 units of O-negative blood this year. It estimates that it will collect blood from 32,000 donors. How great is the risk that the Tennessee Red Cross will fall short of meeting its need? We've just learned how to calculate such probabilities. We can use the Binomial model with  $n = 32,000$  and  $p = 0.06$ . The probability of getting *exactly* 1850 units of O-negative blood from 32,000 donors is  $\binom{32000}{1850} \times 0.06^{1850} \times 0.94^{30150}$ . No calculator on earth can calculate that first term (it has more than 100,000 digits).<sup>1</sup> And that's just the beginning. The problem said *at least* 1850, so we have to do it again for 1851, for 1852, and all the way up to 32,000. No thanks.

When we're dealing with a large number of trials like this, making direct calculations of the probabilities becomes tedious (or outright impossible). Here an old friend—the Normal model—comes to the rescue.

The Binomial model has mean  $np = 1920$  and standard deviation  $\sqrt{npq} \approx 42.48$ . We could try approximating its distribution with a Normal model, using the same mean and standard deviation. Remarkably enough, that turns out to be a very good approximation. (We'll see why in the next chapter.) With that approximation, we can find the *probability*:

$$P(X < 1850) = P\left(z < \frac{1850 - 1920}{42.48}\right) \approx P(z < -1.65) \approx 0.05$$

There seems to be about a 5% chance that this Red Cross chapter will run short of O-negative blood.

Can we always use a Normal model to make estimates of Binomial probabilities? No. Consider the Tiger Woods situation—pictures in 20% of the cereal boxes. If we buy five boxes, the actual Binomial probabilities that we get 0, 1, 2, 3, 4, or 5 pictures of Tiger are 33%, 41%, 20%, 5%, 1%, and 0.03%, respectively. The first histogram shows that this probability model is skewed. That makes it clear that we should not try to estimate these probabilities by using a Normal model.

Now suppose we open 50 boxes of this cereal and count the number of Tiger Woods pictures we find. The second histogram shows this probability model. It is centered at  $np = 50(0.2) = 10$  pictures, as expected, and it appears to be fairly symmetric around that center. Let's have a closer look.

The third histogram again shows Binom(50, 0.2), this time magnified somewhat and centered at the expected value of 10 pictures of Tiger. It looks close to Normal, for sure. With this larger sample size, it appears that a Normal model might be a useful approximation.

A Normal model, then, is a close enough approximation only for a large enough number of trials. And what we mean by "large enough" depends on the probability of success. We'd need a larger sample if the probability of success were very low (or very high). It turns out that a Normal model works pretty well if we expect to see at least 10 successes and 10 failures. That is, we check the **Success/Failure Condition**.

**The Success/ Failure Condition:** A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures:

$$np \geq 10 \text{ and } nq \geq 10.$$

<sup>1</sup> If your calculator *can* find Binom(32000,0.06), then it's smart enough to use an approximation. Read on to see how you can, too.

## MATH BOX

It's easy to see where the magic number 10 comes from. You just need to remember how Normal models work. The problem is that a Normal model extends infinitely in both directions. But a Binomial model must have between 0 and  $n$  successes, so if we use a Normal to approximate a Binomial, we have to cut off its tails. That's not very important if the center of the Normal model is so far from 0 and  $n$  that the lost tails have only a negligible area. More than three standard deviations should do it, because a Normal model has little probability past that.

So the mean needs to be at least 3 standard deviations from 0 and at least 3 standard deviations from  $n$ . Let's look at the 0 end.

We require:	$\mu - 3\sigma > 0$
Or in other words:	$\mu > 3\sigma$
For a Binomial, that's:	$np > 3\sqrt{npq}$
Squaring yields:	$n^2p^2 > 9npq$
Now simplify:	$np > 9q$
Since $q \leq 1$ , we can require:	$np > 9$

For simplicity, we usually require that  $np$  (and  $nq$  for the other tail) be at least 10 to use the Normal approximation, the Success/Failure Condition.<sup>2</sup>

## FOR EXAMPLE

## Spam and the Normal approximation to the Binomial

**Recap:** The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Recently, you installed a spam filter. You observe that over the past week it okayed only 151 of 1422 e-mails you received, classifying the rest as junk. Should you worry that the filtering is too aggressive?

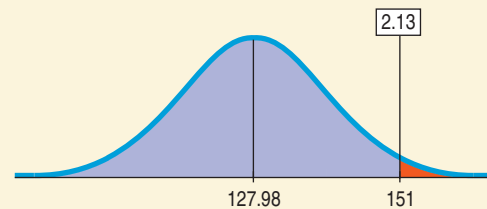
**Question:** What's the probability that no more than 151 of 1422 e-mails is a real message?

I assume that messages arrive randomly and independently, with a probability of success (a real message)  $p = 0.09$ . The model  $\text{Binom}(1422, 0.09)$  applies, but will be hard to work with. Checking conditions for the Normal approximation, I see that:

- ✓ These messages represent less than 10% of all e-mail traffic.
- ✓ I expect  $np = (1422)(0.09) = 127.98$  real messages and  $nq = (1422)(0.91) = 1294.02$  spam messages, both far greater than 10.

It's okay to approximate this binomial probability by using a Normal model.

$$\begin{aligned}\mu &= np = 1422(0.09) = 127.98 \\ \sigma &= \sqrt{npq} = \sqrt{1422(0.09)(0.91)} \approx 10.79 \\ P(x \leq 151) &= P\left(z \leq \frac{151 - 127.98}{10.79}\right) \\ &= P(z \leq 2.13) \\ &= 0.9834\end{aligned}$$



Among my 1422 e-mails, there's over a 98% chance that no more than 151 of them were real messages, so the filter may be working properly.

<sup>2</sup> Looking at the final step, we see that we need  $np > 9$  in the worst case, when  $q$  (or  $p$ ) is near 1, making the Binomial model quite skewed. When  $q$  and  $p$  are near 0.5—say between 0.4 and 0.6—the Binomial model is nearly symmetric and  $np > 5$  ought to be safe enough. Although we'll always check for 10 expected successes and failures, keep in mind that for values of  $p$  near 0.5, we can be somewhat more forgiving.

## Continuous Random Variables

There's a problem with approximating a Binomial model with a Normal model. The Binomial is discrete, giving probabilities for specific counts, but the Normal models a **continuous** random variable that can take on *any value*. For continuous random variables, we can no longer list all the possible outcomes and their probabilities, as we could for discrete random variables.<sup>3</sup>

As we saw in the previous chapter, models for continuous random variables give probabilities for *intervals* of values. So, when we use the Normal model, we no longer calculate the probability that the random variable equals a *particular* value, but only that it lies *between* two values. We won't calculate the probability of getting exactly 1850 units of blood, but we have no problem approximating the probability of getting 1850 *or more*, which was, after all, what we really wanted.<sup>4</sup>



### JUST CHECKING

As we noted a few chapters ago, the Pew Research Center ([www.pewresearch.org](http://www.pewresearch.org)) reports that they are actually able to contact only 76% of the randomly selected households drawn for a telephone survey.

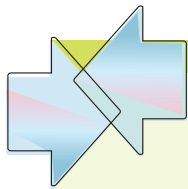
1. Explain why these phone calls can be considered Bernoulli trials.
2. Which of the models of this chapter (Geometric, Binomial, Normal) would you use to model the number of successful contacts from a list of 1000 sampled households? Explain.
3. Pew further reports that even after they contacted a household, only 38% agree to be interviewed, so the probability of getting a completed interview for a randomly selected household is only 0.29. Which of the models of this chapter would you use to model the number of households Pew has to call before they get the first completed interview?

### WHAT CAN GO WRONG?

- ▶ **Be sure you have Bernoulli trials.** Be sure to check the requirements first: two possible outcomes per trial ("success" and "failure"), a constant probability of success, and independence. Remember to check the 10% Condition when sampling without replacement.
- ▶ **Don't confuse Geometric and Binomial models.** Both involve Bernoulli trials, but the issues are different. If you are repeating trials until your first success, that's a Geometric probability. You don't know in advance how many trials you'll need—theoretically, it could take forever. If you are counting the number of successes in a specified number of trials, that's a Binomial probability.
- ▶ **Don't use the Normal approximation with small  $n$ .** To use a Normal approximation in place of a Binomial model, there must be at least 10 expected successes and 10 expected failures.

<sup>3</sup> In fact, some people use an adjustment called the "continuity correction" to help with this problem. It's related to the suggestion we make in the next footnote and is discussed in more advanced textbooks.

<sup>4</sup> If we really had been interested in a single value, we might have approximated it by finding the probability of getting between 1849.5 and 1850.5 units of blood.



## CONNECTIONS

This chapter builds on what we know about random variables. We now have two more probability models to join the Normal model.

There are a number of “forward” connections from this chapter. We’ll see the **10% Condition** and the **Success/Failure Condition** often. And the facts about the Binomial distribution can help explain how proportions behave, as we’ll see in the next chapter.



## WHAT HAVE WE LEARNED?

We’ve learned that Bernoulli trials show up in lots of places. Depending on the random variable of interest, we can use one of three models to estimate probabilities for Bernoulli trials:

- ▶ a Geometric model when we’re interested in the number of Bernoulli trials until the next success;
- ▶ a Binomial model when we’re interested in the number of successes in a certain number of Bernoulli trials;
- ▶ a Normal model to approximate a Binomial model when we expect at least 10 successes and 10 failures.

### Terms

Bernoulli trials, if . . .

388. 1. there are two possible outcomes.  
2. the probability of success is constant.  
3. the trials are independent.

Geometric probability model

389. A Geometric model is appropriate for a random variable that counts the number of Bernoulli trials until the first success.

Binomial probability model

393. A Binomial model is appropriate for a random variable that counts the number of successes in a fixed number of Bernoulli trials.

10% Condition

391. When sampling without replacement, trials are not independent. It’s still okay to proceed as long as the sample is smaller than 10% of the population.

Success/Failure Condition

397. For a Normal model to be a good approximation of a Binomial model, we must expect at least 10 successes and 10 failures. That is,  $np \geq 10$  and  $nq \geq 10$ .

### Skills

THINK

- ▶ Know how to tell if a situation involves Bernoulli trials.
- ▶ Be able to choose whether to use a Geometric or a Binomial model for a random variable involving Bernoulli trials.

SHOW

- ▶ Know the appropriate conditions for using a Geometric, Binomial, or Normal model.
- ▶ Know how to find the expected value of a Geometric model.
- ▶ Be able to calculate Geometric probabilities.
- ▶ Know how to find the mean and standard deviation of a Binomial model.

TELL

- ▶ Be able to calculate Binomial probabilities, perhaps approximating with a Normal model.
- ▶ Be able to interpret means, standard deviations, and probabilities in the Bernoulli trial context.

## THE BINOMIAL AND THE GEOMETRIC ON THE COMPUTER

Most statistics packages offer functions that compute Binomial probabilities, and many offer functions for Geometric probabilities as well. Some technology solutions automatically use the Normal approximation for the Binomial when the exact calculations become unmanageable.

### EXERCISES

- Bernoulli.** Do these situations involve Bernoulli trials? Explain.

  - We roll 50 dice to find the distribution of the number of spots on the faces.
  - How likely is it that in a group of 120 the majority may have Type A blood, given that Type A is found in 43% of the population?
  - We deal 7 cards from a deck and get all hearts. How likely is that?
  - We wish to predict the outcome of a vote on the school budget, and poll 500 of the 3000 likely voters to see how many favor the proposed budget.
  - A company realizes that about 10% of its packages are not being sealed properly. In a case of 24, is it likely that more than 3 are unsealed?
- Bernoulli 2.** Do these situations involve Bernoulli trials? Explain.

  - You are rolling 5 dice and need to get at least two 6's to win the game.
  - We record the distribution of eye colors found in a group of 500 people.
  - A manufacturer recalls a doll because about 3% have buttons that are not properly attached. Customers return 37 of these dolls to the local toy store. Is the manufacturer likely to find any dangerous buttons?
  - A city council of 11 Republicans and 8 Democrats picks a committee of 4 at random. What's the probability they choose all Democrats?
  - A 2002 Rutgers University study found that 74% of high school students have cheated on a test at least once. Your local high school principal conducts a survey in homerooms and gets responses that admit to cheating from 322 of the 481 students.
- Simulating the model.** Think about the Tiger Woods picture search again. You are opening boxes of cereal one at a time looking for his picture, which is in 20% of the boxes. You want to know how many boxes you might have to open in order to find Tiger.

  - Describe how you would simulate the search for Tiger using random numbers.
  - Run at least 30 trials.
  - Based on your simulation, estimate the probabilities that you might find your first picture of Tiger in the first box, the second, etc.
  - Calculate the actual probability model.
  - Compare the distribution of outcomes in your simulation to the probability model.
- Simulation II.** You are one space short of winning a child's board game and must roll a 1 on a die to claim victory. You want to know how many rolls it might take.

  - Describe how you would simulate rolling the die until you get a 1.
  - Run at least 30 trials.
  - Based on your simulation, estimate the probabilities that you might win on the first roll, the second, the third, etc.
  - Calculate the actual probability model.
  - Compare the distribution of outcomes in your simulation to the probability model.
- Tiger again.** Let's take one last look at the Tiger Woods picture search. You know his picture is in 20% of the cereal boxes. You buy five boxes to see how many pictures of Tiger you might get.

  - Describe how you would simulate the number of pictures of Tiger you might find in five boxes of cereal.
  - Run at least 30 trials.
  - Based on your simulation, estimate the probabilities that you get no pictures of Tiger, 1 picture, 2 pictures, etc.
  - Find the actual probability model.
  - Compare the distribution of outcomes in your simulation to the probability model.
- Seatbelts.** Suppose 75% of all drivers always wear their seatbelts. Let's investigate how many of the drivers might be belted among five cars waiting at a traffic light.

  - Describe how you would simulate the number of seatbelt-wearing drivers among the five cars.
  - Run at least 30 trials.
  - Based on your simulation, estimate the probabilities there are no belted drivers, exactly one, two, etc.
  - Find the actual probability model.
  - Compare the distribution of outcomes in your simulation to the probability model.
- On time.** A Department of Transportation report about air travel found that, nationwide, 76% of all flights are on time. Suppose you are at the airport and your flight is one of 50 scheduled to take off in the next two hours. Can you consider these departures to be Bernoulli trials? Explain.



8. **Lost luggage.** A Department of Transportation report about air travel found that airlines misplace about 5 bags per 1000 passengers. Suppose you are traveling with a group of people who have checked 22 pieces of luggage on your flight. Can you consider the fate of these bags to be Bernoulli trials? Explain.
9. **Hoops.** A basketball player has made 80% of his foul shots during the season. Assuming the shots are independent, find the probability that in tonight's game he
- misses for the first time on his fifth attempt.
  - makes his first basket on his fourth shot.
  - makes his first basket on one of his first 3 shots.
10. **Chips.** Suppose a computer chip manufacturer rejects 2% of the chips produced because they fail presale testing.
- What's the probability that the fifth chip you test is the first bad one you find?
  - What's the probability you find a bad one within the first 10 you examine?
11. **More hoops.** For the basketball player in Exercise 9, what's the expected number of shots until he misses?
12. **Chips ahoy.** For the computer chips described in Exercise 10, how many do you expect to test before finding a bad one?
13. **Customer center operator.** Raaj works at the customer service call center of a major credit card bank. Cardholders call for a variety of reasons, but regardless of their reason for calling, if they hold a platinum card, Raaj is instructed to offer them a double-miles promotion. About 10% of all cardholders hold platinum cards, and about 50% of those will take the double-miles promotion. On average, how many calls will Raaj have to take before finding the first cardholder to take the double-miles promotion?
14. **Cold calls.** Justine works for an organization committed to raising money for Alzheimer's research. From past experience, the organization knows that about 20% of all potential donors will agree to give something if contacted by phone. They also know that of all people donating, about 5% will give \$100 or more. On average, how many potential donors will she have to contact until she gets her first \$100 donor?
15. **Blood.** Only 4% of people have Type AB blood.
- On average, how many donors must be checked to find someone with Type AB blood?
  - What's the probability that there is a Type AB donor among the first 5 people checked?
  - What's the probability that the first Type AB donor will be found among the first 6 people?
  - What's the probability that we won't find a Type AB donor before the 10th person?
16. **Colorblindness.** About 8% of males are colorblind. A researcher needs some colorblind subjects for an experiment and begins checking potential subjects.
- On average, how many men should the researcher expect to check to find one who is colorblind?
  - What's the probability that she won't find anyone colorblind among the first 4 men she checks?
  - What's the probability that the first colorblind man found will be the sixth person checked?
  - What's the probability that she finds someone who is colorblind before checking the 10th man?
17. **Lefties.** Assume that 13% of people are left-handed. If we select 5 people at random, find the probability of each outcome described below.
- The first lefty is the fifth person chosen.
  - There are some lefties among the 5 people.
  - The first lefty is the second or third person.
  - There are exactly 3 lefties in the group.
  - There are at least 3 lefties in the group.
  - There are no more than 3 lefties in the group.
18. **Arrows.** An Olympic archer is able to hit the bull's-eye 80% of the time. Assume each shot is independent of the others. If she shoots 6 arrows, what's the probability of each of the following results?
- Her first bull's-eye comes on the third arrow.
  - She misses the bull's-eye at least once.
  - Her first bull's-eye comes on the fourth or fifth arrow.
  - She gets exactly 4 bull's-eyes.
  - She gets at least 4 bull's-eyes.
  - She gets at most 4 bull's-eyes.
19. **Lefties redux.** Consider our group of 5 people from Exercise 17.
- How many lefties do you expect?
  - With what standard deviation?
  - If we keep picking people until we find a lefty, how long do you expect it will take?
20. **More arrows.** Consider our archer from Exercise 18.
- How many bull's-eyes do you expect her to get?
  - With what standard deviation?
  - If she keeps shooting arrows until she hits the bull's-eye, how long do you expect it will take?
21. **Still more lefties.** Suppose we choose 12 people instead of the 5 chosen in Exercise 17.
- Find the mean and standard deviation of the number of right-handers in the group.
  - What's the probability that
    - they're not all right-handed?
    - there are no more than 10 righties?
    - there are exactly 6 of each?
    - the majority is right-handed?
22. **Still more arrows.** Suppose our archer from Exercise 18 shoots 10 arrows.
- Find the mean and standard deviation of the number of bull's-eyes she may get.
  - What's the probability that
    - she never misses?
    - there are no more than 8 bull's-eyes?
    - there are exactly 8 bull's-eyes?
    - she hits the bull's-eye more often than she misses?
23. **Vision.** It is generally believed that nearsightedness affects about 12% of all children. A school district tests the vision of 169 incoming kindergarten children. How many would you expect to be nearsighted? With what standard deviation?

24. **International students.** At a certain college, 6% of all students come from outside the United States. Incoming students there are assigned at random to freshman dorms, where students live in residential clusters of 40 freshmen sharing a common lounge area. How many international students would you expect to find in a typical cluster? With what standard deviation?
25. **Tennis, anyone?** A certain tennis player makes a successful first serve 70% of the time. Assume that each serve is independent of the others. If she serves 6 times, what's the probability she gets
- all 6 serves in?
  - exactly 4 serves in?
  - at least 4 serves in?
  - no more than 4 serves in?
26. **Frogs.** A wildlife biologist examines frogs for a genetic trait he suspects may be linked to sensitivity to industrial toxins in the environment. Previous research had established that this trait is usually found in 1 of every 8 frogs. He collects and examines a dozen frogs. If the frequency of the trait has not changed, what's the probability he finds the trait in
- none of the 12 frogs?
  - at least 2 frogs?
  - 3 or 4 frogs?
  - no more than 4 frogs?
27. **And more tennis.** Suppose the tennis player in Exercise 25 serves 80 times in a match.
- What are the mean and standard deviation of the number of good first serves expected?
  - Verify that you can use a Normal model to approximate the distribution of the number of good first serves.
  - Use the 68–95–99.7 Rule to describe this distribution.
  - What's the probability she makes at least 65 first serves?
28. **More arrows.** The archer in Exercise 18 will be shooting 200 arrows in a large competition.
- What are the mean and standard deviation of the number of bull's-eyes she might get?
  - Is a Normal model appropriate here? Explain.
  - Use the 68–95–99.7 Rule to describe the distribution of the number of bull's-eyes she may get.
  - Would you be surprised if she made only 140 bull's-eyes? Explain.
29. **Apples.** An orchard owner knows that he'll have to use about 6% of the apples he harvests for cider because they will have bruises or blemishes. He expects a tree to produce about 300 apples.
- Describe an appropriate model for the number of cider apples that may come from that tree. Justify your model.
  - Find the probability there will be no more than a dozen cider apples.
  - Is it likely there will be more than 50 cider apples? Explain.
30. **Frogs, part II.** Based on concerns raised by his preliminary research, the biologist in Exercise 26 decides to collect and examine 150 frogs.
- Assuming the frequency of the trait is still 1 in 8, determine the mean and standard deviation of the number of frogs with the trait he should expect to find in his sample.
  - Verify that he can use a Normal model to approximate the distribution of the number of frogs with the trait.
  - He found the trait in 22 of his frogs. Do you think this proves that the trait has become more common? Explain.
31. **Lefties again.** A lecture hall has 200 seats with folding arm tablets, 30 of which are designed for left-handers. The typical size of classes that meet there is 188, and we can assume that about 13% of students are left-handed. What's the probability that a right-handed student in one of these classes is forced to use a lefty arm tablet?
32. **No-shows.** An airline, believing that 5% of passengers fail to show up for flights, overbooks (sells more tickets than there are seats). Suppose a plane will hold 265 passengers, and the airline sells 275 tickets. What's the probability the airline will not have enough seats, so someone gets bumped?
33. **Annoying phone calls.** A newly hired telemarketer is told he will probably make a sale on about 12% of his phone calls. The first week he called 200 people, but only made 10 sales. Should he suspect he was misled about the true success rate? Explain.
34. **The euro.** Shortly after the introduction of the euro coin in Belgium, newspapers around the world published articles claiming the coin is biased. The stories were based on reports that someone had spun the coin 250 times and gotten 140 heads—that's 56% heads. Do you think this is evidence that spinning a euro is unfair? Explain.
35. **Seatbelts II.** Police estimate that 80% of drivers now wear their seatbelts. They set up a safety roadblock, stopping cars to check for seatbelt use.
- How many cars do they expect to stop before finding a driver whose seatbelt is not buckled?
  - What's the probability that the first unbelted driver is in the 6th car stopped?
  - What's the probability that the first 10 drivers are all wearing their seatbelts?
  - If they stop 30 cars during the first hour, find the mean and standard deviation of the number of drivers expected to be wearing seatbelts.
  - If they stop 120 cars during this safety check, what's the probability they find at least 20 drivers not wearing their seatbelts?
36. **Rickets.** Vitamin D is essential for strong, healthy bones. Our bodies produce vitamin D naturally when sunlight falls upon the skin, or it can be taken as a dietary supplement. Although the bone disease rickets was largely eliminated in England during the 1950s, some people there are concerned that this generation of children is at increased risk because they are more likely to watch TV or play computer games than spend time outdoors. Recent research indicated that about 20% of British children are deficient in vitamin D. Suppose doctors test a group of elementary school children.

- a) What's the probability that the first vitamin D-deficient child is the 8th one tested?
- b) What's the probability that the first 10 children tested are all okay?
- c) How many kids do they expect to test before finding one who has this vitamin deficiency?
- d) They will test 50 students at the third-grade level. Find the mean and standard deviation of the number who may be deficient in vitamin D.
- e) If they test 320 children at this school, what's the probability that no more than 50 of them have the vitamin deficiency?
37. **ESP.** Scientists wish to test the mind-reading ability of a person who claims to "have ESP." They use five cards with different and distinctive symbols (square, circle, triangle, line, squiggle). Someone picks a card at random and thinks about the symbol. The "mind reader" must correctly identify which symbol was on the card. If the test consists of 100 trials, how many would this person need to get right in order to convince you that ESP may actually exist? Explain.
38. **True-False.** A true-false test consists of 50 questions. How many does a student have to get right to convince you that he is not merely guessing? Explain.
39. **Hot hand.** A basketball player who ordinarily makes about 55% of his free throw shots has made 4 in a row. Is this evidence that he has a "hot hand" tonight? That is, is this streak so unusual that it means the probability he makes a shot must have changed? Explain.
40. **New bow.** Our archer in Exercise 18 purchases a new bow, hoping that it will improve her success rate to more than 80% bull's-eyes. She is delighted when she first tests

her new bow and hits 6 consecutive bull's-eyes. Do you think this is compelling evidence that the new bow is better? In other words, is a streak like this unusual for her? Explain.

41. **Hotter hand.** Our basketball player in Exercise 39 has new sneakers, which he thinks improve his game. Over his past 40 shots, he's made 32—much better than the 55% he usually shoots. Do you think his chances of making a shot really increased? In other words, is making at least 32 of 40 shots really unusual for him? (Do you think it's his sneakers?)
42. **New bow, again.** The archer in Exercise 40 continues shooting arrows, ending up with 45 bull's-eyes in 50 shots. Now are you convinced that the new bow is better? Explain.



### JUST CHECKING Answers

1. There are two outcomes (contact, no contact), the probability of contact is 0.76, and random calls should be independent.
2. Binomial, with  $n = 1000$  and  $p = 0.76$ . For actual calculations, we could approximate using a Normal model with  $\mu = np = 1000(0.76) = 760$  and
 
$$\sigma = \sqrt{npq} = \sqrt{1000(0.76)(0.24)} \approx 13.5.$$
3. Geometric, with  $p = 0.29$ .

## REVIEW OF PART IV

## Randomness and Probability

## Quick Review

Here's a brief summary of the key concepts and skills in probability and probability modeling:

- ▶ The Law of Large Numbers says that the more times we try something, the closer the results will come to theoretical perfection.
  - Don't mistakenly misinterpret the Law of Large Numbers as the "Law of Averages." There's no such thing.
- ▶ Basic rules of probability can handle most situations:
  - To find the probability that an event OR another event happens, add their probabilities and subtract the probability that both happen.
  - To find the probability that an event AND another independent event both happen, multiply probabilities.
  - Conditional probabilities tell you how likely one event is to happen, knowing that another event has happened.
  - Mutually exclusive events (also called "disjoint") cannot both happen at the same time.
  - Two events are independent if the occurrence of one doesn't change the probability that the other happens.
- ▶ A probability model for a random variable describes the theoretical distribution of outcomes.
  - The mean of a random variable is its expected value.
  - For sums or differences of independent random variables, variances add.
  - To estimate probabilities involving quantitative variables, you may be able to use a Normal model—but only if the distribution of the variable is unimodal and symmetric.
  - To estimate the probability you'll get your first success on a certain trial, use a Geometric model.
  - To estimate the probability you'll get a certain number of successes in a specified number of independent trials, use a Binomial model.

Ready? Here are some opportunities to check your understanding of these ideas.

## REVIEW EXERCISES

1. **Quality control.** A consumer organization estimates that 29% of new cars have a cosmetic defect, such as a scratch or a dent, when they are delivered to car dealers. This same organization believes that 7% have a functional defect—something that does not work properly—and that 2% of new cars have both kinds of problems.
  - a) If you buy a new car, what's the probability that it has some kind of defect?
  - b) What's the probability it has a cosmetic defect but no functional defect?
  - c) If you notice a dent on a new car, what's the probability it has a functional defect?
  - d) Are the two kinds of defects disjoint events? Explain.
  - e) Do you think the two kinds of defects are independent events? Explain.
2. **Workers.** A company's human resources officer reports a breakdown of employees by job type and sex shown in the table.
 

		Sex	
		Male	Female
Job Type	Management	7	6
	Supervision	8	12
	Production	45	72

  - a) What's the probability that a worker selected at random is
    - i) female?
    - ii) female or a production worker?
    - iii) female, if the person works in production?
    - iv) a production worker, if the person is female?
  - b) Do these data suggest that job type is independent of being male or female? Explain.
3. **Airfares.** Each year a company must send 3 officials to a meeting in China and 5 officials to a meeting in France. Airline ticket prices vary from time to time, but the company purchases all tickets for a country at the same price. Past experience has shown that tickets to China have a mean price of \$1000, with a standard deviation of \$150, while the mean airfare to France is \$500, with a standard deviation of \$100.
  - a) Define random variables and use them to express the total amount the company will have to spend to send these delegations to the two meetings.
  - b) Find the mean and standard deviation of this total cost.
  - c) Find the mean and standard deviation of the difference in price of a ticket to China and a ticket to France.
  - d) Do you need to make any assumptions in calculating these means? How about the standard deviations?

4. **Bipolar.** Psychiatrists estimate that about 1 in 100 adults suffers from bipolar disorder. What's the probability that in a city of 10,000 there are more than 200 people with this condition? Be sure to verify that a Normal model can be used here.
5. **A game.** To play a game, you must pay \$5 for each play. There is a 10% chance you will win \$5, a 40% chance you will win \$7, and a 50% chance you will win only \$3.
- What are the mean and standard deviation of your net winnings?
  - You play twice. Assuming the plays are independent events, what are the mean and standard deviation of your total winnings?
6. **Emergency switch.** Safety engineers must determine whether industrial workers can operate a machine's emergency shutoff device. Among a group of test subjects, 66% were successful with their left hands, 82% with their right hands, and 51% with either hand.
- What percent of these workers could not operate the switch with either hand?
  - Are success with right and left hands independent events? Explain.
  - Are success with right and left hands mutually exclusive? Explain.
7. **Twins.** In the United States, the probability of having twins (usually about 1 in 90 births) rises to about 1 in 10 for women who have been taking the fertility drug Clomid. Among a group of 10 pregnant women, what's the probability that
- at least one will have twins if none were taking a fertility drug?
  - at least one will have twins if all were taking Clomid?
  - at least one will have twins if half were taking Clomid?
8. **Deductible.** A car owner may buy insurance that will pay the full price of repairing the car after an at-fault accident, or save \$12 a year by getting a policy with a \$500 deductible. Her insurance company says that about 0.5% of drivers in her area have an at-fault auto accident during any given year. Based on this information, should she buy the policy with the deductible or not? How does the value of her car influence this decision?
9. **More twins.** A group of 5 women became pregnant while undergoing fertility treatments with the drug Clomid, discussed in Exercise 7. What's the probability that
- none will have twins?
  - exactly 1 will have twins?
  - at least 3 will have twins?
10. **At fault.** The car insurance company in Exercise 8 believes that about 0.5% of drivers have an at-fault accident during a given year. Suppose the company insures 1355 drivers in that city.
- What are the mean and standard deviation of the number who may have at-fault accidents?
  - Can you describe the distribution of these accidents with a Normal model? Explain.
11. **Twins, part III.** At a large fertility clinic, 152 women became pregnant while taking Clomid. (See Exercise 7.)
- What are the mean and standard deviation of the number of twin births we might expect?
  - Can we use a Normal model in this situation? Explain.
  - What's the probability that no more than 10 of the women have twins?
12. **Child's play.** In a board game you determine the number of spaces you may move by spinning a spinner and rolling a die. The spinner has three regions: Half of the spinner is marked "5," and the other half is equally divided between "10" and "20." The six faces of the die show 0, 0, 1, 2, 3, and 4 spots. When it's your turn, you spin and roll, adding the numbers together to determine how far you may move.
- Create a probability model for the outcome on the spinner.
  - Find the mean and standard deviation of the spinner results.
  - Create a probability model for the outcome on the die.
  - Find the mean and standard deviation of the die results.
  - Find the mean and standard deviation of the number of spaces you get to move.
13. **Language.** Neurological research has shown that in about 80% of people, language abilities reside in the brain's left side. Another 10% display right-brain language centers, and the remaining 10% have two-sided language control. (The latter two groups are mainly left-handers; *Science News*, 161 no. 24 [2002].)
- Assume that a freshman composition class contains 25 randomly selected people. What's the probability that no more than 15 of them have left-brain language control?
  - In a randomly chosen group of 5 of these students, what's the probability that no one has two-sided language control?
  - In the entire freshman class of 1200 students, how many would you expect to find of each type?
  - What are the mean and standard deviation of the number of these freshmen who might be right-brained in language abilities?
  - If an assumption of Normality is justified, use the 68–95–99.7 Rule to describe how many students in the freshman class might have right-brain language control.
14. **Play again.** If you land in a "penalty zone" on the game board described in Exercise 12, your move will be determined by subtracting the roll of the die from the result on the spinner. Now what are the mean and standard deviation of the number of spots you may move?
15. **Beanstalks.** In some cities tall people who want to meet and socialize with other tall people can join Beanstalk Clubs. To qualify, a man must be over 6'2" tall, and a woman over 5'10". According to the National Health Survey, heights of adults may have a Normal model with mean heights of 69.1" for men and 64.0" for women. The respective standard deviations are 2.8" and 2.5".

- a) You're probably not surprised to learn that men are generally taller than women, but what does the greater standard deviation for men's heights indicate?
- b) Are men or women more likely to qualify for Beanstalk membership?
- c) Beanstalk members believe that height is an important factor when people select their spouses. To investigate, we select at random a married man and, independently, a married woman. Define two random variables, and use them to express how many inches taller the man is than the woman.
- d) What's the mean of this difference?
- e) What's the standard deviation of this difference?
- f) What's the probability that the man is taller than the woman (that the difference in heights is greater than 0)?
- g) Suppose a survey of married couples reveals that 92% of the husbands were taller than their wives. Based on your answer to part f, do you believe that people's choice of spouses is independent of height? Explain.
16. **Stocks.** Since the stock market began in 1872, stock prices have risen in about 73% of the years. Assuming that market performance is independent from year to year, what's the probability that
- the market will rise for 3 consecutive years?
  - the market will rise 3 years out of the next 5?
  - the market will fall during at least 1 of the next 5 years?
  - the market will rise during a majority of years over the next decade?
17. **Multiple choice.** A multiple choice test has 50 questions, with 4 answer choices each. You must get at least 30 correct to pass the test, and the questions are very difficult.
- Are you likely to be able to pass by guessing on every question? Explain.
  - Suppose, after studying for a while, you believe you have raised your chances of getting each question right to 70%. How likely are you to pass now?
  - Assuming you are operating at the 70% level and the instructor arranges questions randomly, what's the probability that the third question is the first one you get right?
18. **Stock strategy.** Many investment advisors argue that after stocks have declined in value for 2 consecutive years, people should invest heavily because the market rarely declines 3 years in a row.
- Since the stock market began in 1872, there have been two consecutive losing years eight times. In six of those cases, the market rose during the following year. Does this confirm the advice?
  - Overall, stocks have risen in value during 95 of the 130 years since the market began in 1872. How is this fact relevant in assessing the statistical reasoning of the advisors?
19. **Insurance.** A 65-year-old woman takes out a \$10,000 term life insurance policy. The company charges an annual premium of \$500. Estimate the company's expected profit on such policies if mortality tables indicate that only 2.6% of women age 65 die within a year.
20. **Teen smoking.** The Centers for Disease Control say that about 30% of high-school students smoke tobacco (down from a high of 38% in 1997). Suppose you randomly select high-school students to survey them on their attitudes toward scenes of smoking in the movies. What's the probability that
- none of the first 4 students you interview is a smoker?
  - the first smoker is the sixth person you choose?
  - there are no more than 2 smokers among 10 people you choose?
21. **Passing stats.** Molly's college offers two sections of Statistics 101. From what she has heard about the two professors listed, Molly estimates that her chances of passing the course are 0.80 if she gets Professor Scedastic and 0.60 if she gets Professor Kurtosis. The registrar uses a lottery to randomly assign the 120 enrolled students based on the number of available seats in each class. There are 70 seats in Professor Scedastic's class and 50 in Professor Kurtosis's class.
- What's the probability that Molly will pass Statistics?
  - At the end of the semester, we find out that Molly failed. What's the probability that she got Professor Kurtosis?
22. **Teen smoking II.** Suppose that, as reported by the Centers for Disease Control, about 30% of high school students smoke tobacco. You randomly select 120 high school students to survey them on their attitudes toward scenes of smoking in the movies.
- What's the expected number of smokers?
  - What's the standard deviation of the number of smokers?
  - The number of smokers among 120 randomly selected students will vary from group to group. Explain why that number can be described with a Normal model.
  - Using the 68–95–99.7 Rule, create and interpret a model for the number of smokers among your group of 120 students.
23. **Random variables.** Given independent random variables with means and standard deviations as shown, find the mean and standard deviation of each of these variables:
- |     | Mean | SD |
|-----|------|----|
| $X$ | 50   | 8  |
| $Y$ | 100  | 6  |
- $X + 50$
  - $10Y$
  - $X + 0.5Y$
  - $X - Y$
  - $X_1 + X_2$
24. **Merger.** Explain why the facts you know about variances of independent random variables might encourage two small insurance companies to merge. (*Hint:* Think about the expected amount and potential variability in payouts for the separate and the merged companies.)
25. **Youth survey.** According to a recent Gallup survey, 93% of teens use the Internet, but there are differences in how teen boys and girls say they use computers. The telephone poll found that 77% of boys had played computer games in the past week, compared with 65% of girls. On the other hand, 76% of girls said they had e-mailed friends in the past week, compared with only 65% of boys.

- a) For boys, the cited percentages are 77% playing computer games and 65% using e-mail. That total is 142%, so there is obviously a mistake in the report. No? Explain.
- b) Based on these results, do you think playing games and using e-mail are mutually exclusive? Explain.
- c) Do you think whether a child e-mails friends is independent of being a boy or a girl? Explain.
- d) Suppose that in fact 93% of the teens in your area do use the Internet. You want to interview a few who do not, so you start contacting teenagers at random. What is the probability that it takes you 5 interviews until you find the first person who does not use the Internet?
26. **Meals.** A college student on a seven-day meal plan reports that the amount of money he spends daily on food varies with a mean of \$13.50 and a standard deviation of \$7.
- a) What are the mean and standard deviation of the amount he might spend in two consecutive days?
- b) What assumption did you make in order to find that standard deviation? Are there any reasons you might question that assumption?
- c) Estimate his average weekly food costs, and the standard deviation.
- d) Do you think it likely he might spend less than \$50 in a week? Explain, including any assumptions you make in your analysis.
27. **Travel to Kyrgyzstan.** Your pocket copy of *Kyrgyzstan on 4237 ± 360 Som a Day* claims that you can expect to spend about 4237 som each day with a standard deviation of 360 som. How well can you estimate your expenses for the trip?
- a) Your budget allows you to spend 90,000 som. To the nearest day, how long can you afford to stay in Kyrgyzstan, on average?
- b) What's the standard deviation of your expenses for a trip of that duration?
- c) You doubt that your total expenses will exceed your expectations by more than two standard deviations. How much extra money should you bring? On average, how much of a "cushion" will you have per day?
28. **Picking melons.** Two stores sell watermelons. At the first store the melons weigh an average of 22 pounds, with a standard deviation of 2.5 pounds. At the second store the melons are smaller, with a mean of 18 pounds and a standard deviation of 2 pounds. You select a melon at random at each store.
- a) What's the mean difference in weights of the melons?
- b) What's the standard deviation of the difference in weights?
- c) If a Normal model can be used to describe the difference in weights, what's the probability that the melon you got at the first store is heavier?
29. **Home, sweet home.** According to the 2000 Census, 66% of U.S. households own the home they live in. A mayoral candidate conducts a survey of 820 randomly selected homes in your city and finds only 523 owned
- by the current residents. The candidate then attacks the incumbent mayor, saying that there is an unusually low level of homeownership in the city. Do you agree? Explain.
30. **Buying melons.** The first store in Exercise 28 sells watermelons for 32 cents a pound. The second store is having a sale on watermelons—only 25 cents a pound. Find the mean and standard deviation of the difference in the price you may pay for melons randomly selected at each store.
31. **Who's the boss?** The 2000 Census revealed that 26% of all firms in the United States are owned by women. You call some firms doing business locally, assuming that the national percentage is true in your area.
- a) What's the probability that the first 3 you call are all owned by women?
- b) What's the probability that none of your first 4 calls finds a firm that is owned by a woman?
- c) Suppose none of your first 5 calls found a firm owned by a woman. What's the probability that your next call does?
32. **Jerseys.** A Statistics professor comes home to find that all four of his children got white team shirts from soccer camp this year. He concludes that this year, unlike other years, the camp must not be using a variety of colors. But then he finds out that in each child's age group there are 4 teams, only 1 of which wears white shirts. Each child just happened to get on the white team at random.
- a) Why was he so surprised? If each age group uses the same 4 colors, what's the probability that all four kids would get the same-color shirt?
- b) What's the probability that all 4 would get white shirts?
- c) We lied. Actually, in the oldest child's group there are 6 teams instead of the 4 teams in each of the other three groups. How does this change the probability you calculated in part b)?



33. **When to stop?** In Exercise 27 of the Review Exercises for Part III, we posed this question:

*You play a game that involves rolling a die. You can roll as many times as you want, and your score is the total for all the rolls. But . . . if you roll a 6, your score is 0 and your turn is over. What might be a good strategy for a game like this?*

You attempted to devise a good strategy by simulating several plays to see what might happen. Let's try calculating a strategy.

- a) On what roll would you expect to get a 6 for the first time?
- b) So, roll *one time less* than that. Assuming all those rolls were not 6's, what's your expected score?
- c) What's the probability that you can roll that many times without getting a 6?
34. **Plan B.** Here's another attempt at developing a good strategy for the dice game in Exercise 33. Instead of stopping after a certain number of rolls, you could decide to stop when your score reaches a certain number of points.
- a) How many points would you expect a roll to *add* to your score?
- b) In terms of your current score, how many points would you expect a roll to *subtract* from your score?
- c) Based on your answers in parts a and b, at what score will another roll "break even"?
- d) Describe the strategy this result suggests.
35. **Technology on campus.** Every 5 years the Conference Board of the Mathematical Sciences surveys college math departments. In 2000 the board reported that 51% of all undergraduates taking Calculus I were in classes that used graphing calculators and 31% were in classes that used computer assignments. Suppose that 16% used both calculators and computers.
- a) What percent used neither kind of technology?
- b) What percent used calculators but not computers?
- c) What percent of the calculator users had computer assignments?
- d) Based on this survey, do calculator and computer use appear to be independent events? Explain.
36. **Dogs.** A census by the county dog control officer found that 18% of homes kept one dog as a pet, 4% had two dogs, and 1% had three or more. If a salesman visits two homes selected at random, what's the probability he encounters
- a) no dogs?
- b) some dogs?
- c) dogs in each home?
- d) more than one dog in each home?
37. **Socks.** In your sock drawer you have 4 blue socks, 5 grey socks, and 3 black ones. Half asleep one morning, you grab 2 socks at random and put them on. Find the probability you end up wearing
- a) 2 blue socks.
- b) no grey socks.
- c) at least 1 black sock.
- d) a green sock.
- e) matching socks.
38. **Coins.** A coin is to be tossed 36 times.
- a) What are the mean and standard deviation of the number of heads?
- b) Suppose the resulting number of heads is unusual, two standard deviations above the mean. How many "extra" heads were observed?
- c) If the coin were tossed 100 times, would you still consider the same number of extra heads unusual? Explain.
- d) In the 100 tosses, how many extra heads would you need to observe in order to say the results were unusual?
- e) Explain how these results refute the "Law of Averages" but confirm the Law of Large Numbers.
39. **The Drake equation.** In 1961 astronomer Frank Drake developed an equation to try to estimate the number of extraterrestrial civilizations in our galaxy that might be able to communicate with us via radio transmissions. Now largely accepted by the scientific community, the Drake equation has helped spur efforts by radio astronomers to search for extraterrestrial intelligence. Here is the equation:
- $$N_C = N \cdot f_p \cdot n_e \cdot f_l \cdot f_i \cdot f_c \cdot f_L$$
- OK, it looks a little messy, but here's what it means:

Factor	What It Represents	Possible Value
$N$	Number of stars in the Milky Way Galaxy	200–400 billion
$f_p$	Probability that a star has planets	20%–50%
$n_e$	Number of planets in a solar system capable of sustaining earth-type life	1? 2?
$f_l$	Probability that life develops on a planet with a suitable environment	1%–100%
$f_i$	Probability that life evolves intelligence	50%?
$f_c$	Probability that intelligent life develops radio communication	10%–20%
$f_L$	Fraction of the planet's life for which the civilization survives	$\frac{1}{1,000,000}$ ?
$N_c$	Number of extraterrestrial civilizations in our galaxy with which we could communicate	?

So, how many ETs are out there? That depends; values chosen for the many factors in the equation depend on ever-evolving scientific knowledge and one's personal guesses. But now, some questions.

- a) What quantity is calculated by the first product,  $N \cdot f_p$ ?
- b) What quantity is calculated by the product,  $N \cdot f_p \cdot n_e \cdot f_l$ ?
- c) What probability is calculated by the product  $f_l \cdot f_i$ ?
- d) Which of the factors in the formula are conditional probabilities? Restate each in a way that makes the condition clear.

*Note:* A quick Internet search will find you a site where you can play with the Drake equation yourself.

40. **Recalls.** In a car rental company's fleet, 70% of the cars are American brands, 20% are Japanese, and the rest are German. The company notes that manufacturers' recalls seem to affect 2% of the American cars, but only 1% of the others.
- a) What's the probability that a randomly chosen car is recalled?
- b) What's the probability that a recalled car is American?



41. **Pregnant?** Suppose that 70% of the women who suspect they may be pregnant and purchase an in-home pregnancy test are actually pregnant. Further suppose that the test is 98% accurate. What's the probability that a woman whose test indicates that she is pregnant actually is?
42. **Door prize.** You are among 100 people attending a charity fundraiser at which a large-screen TV will be given away as a door prize. To determine who wins, 99 white balls and 1 red ball have been placed in a box and thoroughly mixed. The guests will line up and, one at a time, pick a ball from the box. Whoever gets the red ball wins the TV, but if the ball is white, it is returned to the box. If none of the 100 guests gets the red ball, the TV will be auctioned off for additional benefit of the charity.
- a) What's the probability that the first person in line wins the TV?
- b) You are the third person in line. What's the probability that you win the TV?
- c) What's the probability that the charity gets to sell the TV because no one wins?
- d) Suppose you get to pick your spot in line. Where would you want to be in order to maximize your chances of winning?
- e) After hearing some protest about the plan, the organizers decide to award the prize by not returning the white balls to the box, thus ensuring that 1 of the 100 people will draw the red ball and win the TV. Now what position in line would you choose in order to maximize your chances?



PART

# From the Data at Hand to the World at Large

## Chapter 18

Sampling Distribution Models

## Chapter 19

Confidence Intervals for Proportions

## Chapter 20

Testing Hypotheses About Proportions

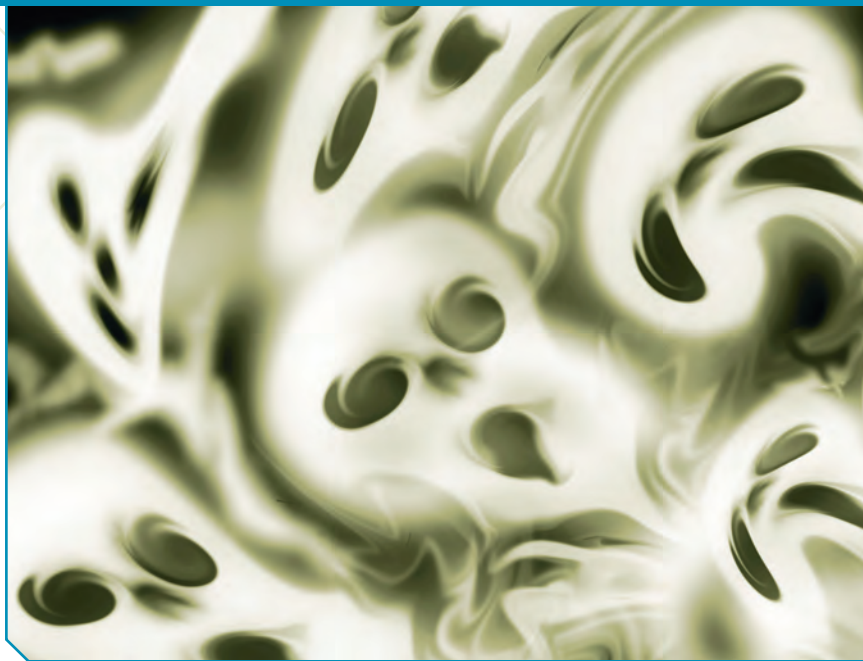
## Chapter 21

More About Tests and Intervals

## Chapter 22

Comparing Two Proportions

# Sampling Distribution Models



<b>WHO</b>	U.S. adults
<b>WHAT</b>	Belief in ghosts
<b>WHEN</b>	November 2005
<b>WHERE</b>	United States
<b>WHY</b>	Public attitudes

In November 2005 the Harris Poll asked 889 U.S. adults, “Do you believe in ghosts?” 40% said they did. At almost the same time, CBS News polled 808 U.S. adults and asked the same question. 48% of their respondents professed a belief in ghosts. Why the difference? This seems like a simple enough question. Should we be surprised to find that we could get proportions this different from properly selected random samples drawn from the same population? You’re probably used to seeing that observations vary, but how much variability among polls should we expect to see?

Why do sample proportions vary at all? How can surveys conducted at essentially the same time by organizations asking the same questions get different results? The answer is at the heart of Statistics. The proportions vary from sample to sample because the samples are composed of different people.

It’s actually pretty easy to predict how much a proportion will vary under circumstances like this. Understanding the variability of our estimates will let us actually use that variability to better understand the world.

## The Central Limit Theorem for Sample Proportions

### Imagine

We see only the sample that we actually drew, but by simulating or modeling, we can *imagine* what we might have seen had we drawn other possible random samples.

We’ve talked about *Think*, *Show*, and *Tell*. Now we have to add *Imagine*. In order to understand the CBS poll, we want to imagine the results from all the random samples of size 808 that CBS News didn’t take. What would the histogram of all the sample proportions look like?

For people’s belief in ghosts, where do you expect the center of that histogram to be? Of course, we don’t *know* the answer to that (and probably never will). But we know that it will be at the true proportion in the population, and we can call that  $p$ . (See the Notation Alert.) For the sake of discussion here, let’s suppose that 45% of all American adults believe in ghosts, so we’ll use  $p = 0.45$ .

How about the *shape* of the histogram? We don’t have to just imagine. We can simulate a bunch of random samples that we didn’t really draw. Here’s a histogram of the proportions saying they believe in ghosts for 2000 simulated independent samples of 808 adults when the true proportion is  $p = 0.45$ .

**A S** **Activity: Sampling Distribution of a Proportion.** You don't have to imagine—you can simulate.

**TI-*n*spire**

**Sample Proportions.** Generate sample after sample to see how the proportions vary.

**NOTATION ALERT:**

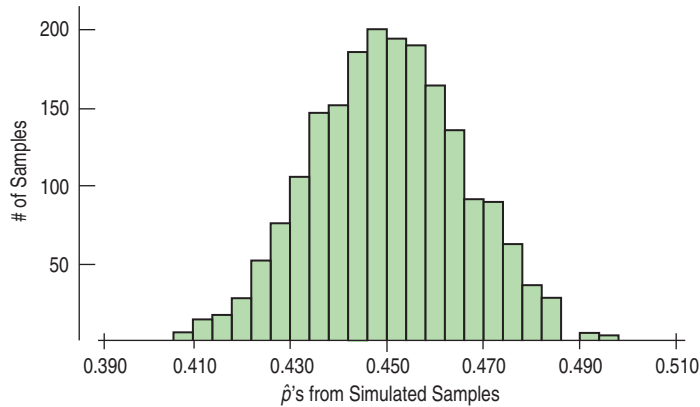
The letter  $p$  is our choice for the *parameter* of the model for proportions. It violates our “Greek letters for parameters” rule, but if we stuck to that, our natural choice would be  $\pi$ . We could use  $\pi$  to be perfectly consistent, but then we'd have to write statements like  $\pi = 0.46$ . That just seems a bit weird to us. After all, we've known that  $\pi = 3.1415926 \dots$  since the Greeks, and it's a hard habit to break.

So, we'll use  $p$  for the model parameter (the probability of a success) and  $\hat{p}$  for the observed proportion in a sample. We'll also use  $q$  for the probability of a failure ( $q = 1 - p$ ) and  $\hat{q}$  for its observed value.

But be careful. We've already used capital  $P$  for a general probability. And we'll soon see another use of  $P$  in the next chapter! There are a lot of  $p$ 's in this course; you'll need to think clearly about the context to keep them straight.



Pierre-Simon Laplace, 1749–1827.



**FIGURE 18.1**

A histogram of sample proportions for 2000 simulated samples of 808 adults drawn from a population with  $p = 0.45$ . The sample proportions vary, but their distribution is centered at the true proportion,  $p$ .

It should be no surprise that we don't get the same proportion for each sample we draw, even though the underlying true value is the same for the population. Each  $\hat{p}$  comes from a different simulated sample. The histogram above is a simulation of what we'd get if we could see *all the proportions from all possible samples*. That distribution has a special name. It is called the **sampling distribution of the proportions**.<sup>1</sup>

Does it surprise you that the histogram is unimodal? Symmetric? That it is centered at  $p$ ? You probably don't find any of this shocking. Does the shape remind you of any model that we've discussed? It's an amazing and fortunate fact that a Normal model is just the right one for the histogram of sample proportions.

As we'll see in a few pages, this fact was proved in 1810 by the great French mathematician Pierre-Simon Laplace as part of a more general result. There is no reason you should guess that the Normal model would be the one we need here,<sup>2</sup> and, indeed, the importance of Laplace's result was not immediately understood by his contemporaries. But (unlike Laplace's contemporaries in 1810) we know how useful the Normal model can be.

Modeling how sample proportions vary from sample to sample is one of the most powerful ideas we'll see in this course. A **sampling distribution model** for how a sample proportion varies from sample to sample allows us to quantify that variation and to talk about how likely it is that we'd observe a sample proportion in any particular interval.

To use a Normal model, we need to specify two parameters: its mean and standard deviation. The center of the histogram is naturally at  $p$ , so we'll put  $\mu$ , the mean of the Normal, at  $p$ .

What about the standard deviation? Usually the mean gives us no information about the standard deviation. Suppose we told you that a batch of bike helmets had a mean diameter of 26 centimeters and asked what the standard deviation was. If you said, "I have no idea," you'd be exactly right. There's no information about  $\sigma$  from knowing the value of  $\mu$ .

But there's a special fact about proportions. With proportions we get something for free. Once we know the mean,  $p$ , we automatically also know the standard deviation. We saw in the last chapter that for a Binomial model the standard deviation of the *number* of successes is  $\sqrt{npq}$ . Now we want the standard deviation

<sup>1</sup> A word of caution. Until now we've been plotting the *distribution of the sample*, a display of the actual data that were collected in that one sample. But now we've plotted the *sampling distribution*; a display of summary statistics ( $\hat{p}$ 's, for example) for many different samples. "Sample distribution" and "sampling distribution" sound a lot alike, but they refer to very different things. (Sorry about that—we didn't make up the terms. It's just the way it is.) And the distinction is critical. Whenever you read or write something about one of these, think very carefully about what the words signify.

<sup>2</sup> Well, the fact that we spent most of Chapter 6 on the Normal model might have been a hint.

of the *proportion* of successes,  $\hat{p}$ . The sample proportion  $\hat{p}$  is the number of successes divided by the number of trials,  $n$ , so the standard deviation is also divided by  $n$ :

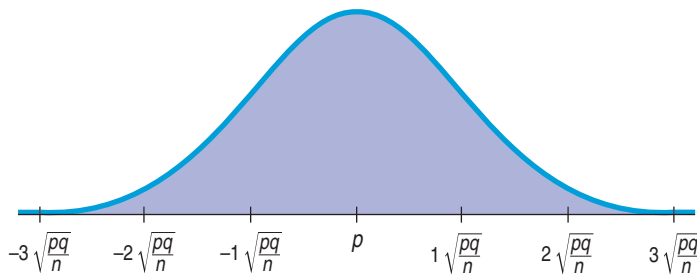
$$\sigma(\hat{p}) = SD(\hat{p}) = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

When we draw simple random samples of  $n$  individuals, the proportions we find will vary from sample to sample. As long as  $n$  is reasonably large,<sup>3</sup> we can model the distribution of these sample proportions with a probability model that is

$$N\left(p, \sqrt{\frac{pq}{n}}\right).$$

**A S** **Simulation: Simulating Sampling Distributions.** Watch the Normal model appear from random proportions.

**FIGURE 18.2**  
A Normal model centered at  $p$  with a standard deviation of  $\sqrt{\frac{pq}{n}}$  is a good model for a collection of proportions found for many random samples of size  $n$  from a population with success probability  $p$ .



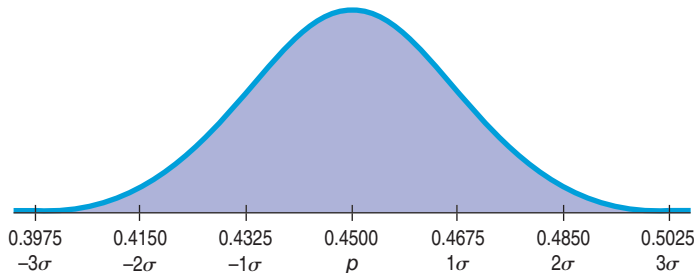
**NOTATION ALERT:**  
In Chapter 8 we introduced  $\hat{y}$  as the predicted value for  $y$ . The “hat” here plays a similar role. It indicates that  $\hat{p}$ —the observed proportion in our data—is our *estimate* of the parameter  $p$ .

Although we’ll never know the true proportion of adults who believe in ghosts, we’re supposing it to be 45%. Once we put the center at  $p = 0.45$ , the standard deviation for the CBS poll is

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.45)(0.55)}{808}} = 0.0175, \text{ or } 1.75\%.$$

Here’s a picture of the Normal model for our simulation histogram:

**FIGURE 18.3**  
Using 0.45 for  $p$  gives this Normal model for Figure 18.1’s histogram of the sample proportions of adults believing in ghosts ( $n = 808$ ).



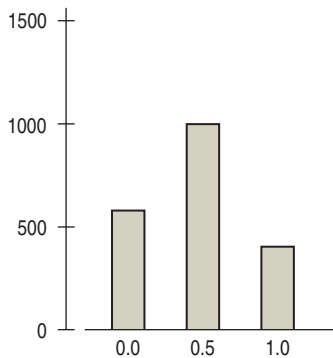
**A S** **Simulation: The Standard Deviation of a Proportion.** Do you believe this formula for standard deviation? Don’t just take our word for it—convince yourself with an experiment.

Because we have a Normal model, we can use the 68–95–99.7 Rule or look up other probabilities using a table or technology. For example, we know that 95% of Normally distributed values are within two standard deviations of the mean, so we should not be surprised if 95% of various polls gave results that were near 45% but varied above and below that by no more than two standard deviations. Since  $2 \times 1.75\% = 3.5\%$ ,<sup>4</sup> we see that the CBS poll estimating belief in ghosts at 48% is *consistent* with our guess of 45%. This is what we mean by **sampling error**. It’s not really an *error* at all, but just *variability* you’d expect to see from one sample to another. A better term would be **sampling variability**.

<sup>3</sup> For smaller  $n$ , we can just use a Binomial model.

<sup>4</sup> The standard deviation is 1.75%. Remember that the standard deviation always has the same units as the data. Here our units are %. But that can be confusing, because the standard deviation is not 1.75% of anything. It is 1.75 percentage points. If that’s confusing, try writing the units as “percentage points” instead of %.

## How Good Is the Normal Model?



**FIGURE 18.4**

Proportions from samples of size 2 can take on only three possible values. A Normal model does not work well.

Stop and think for a minute about what we've just said. It's a remarkable claim. We've said that if we draw repeated random samples of the same size,  $n$ , from some population and measure the proportion,  $\hat{p}$ , we see in each sample, then the collection of these proportions will pile up around the underlying population proportion,  $p$ , and that a histogram of the sample proportions can be modeled well by a Normal model.

There must be a catch. Suppose the samples were of size 2, for example. Then the only possible proportion values would be 0, 0.5, and 1. There's no way the histogram could ever look like a Normal model with only three possible values for the variable.

Well, there *is* a catch. The claim is only approximately true. (But, that's OK. After all, models are only supposed to be approximately true.) And the model becomes a better and better representation of the distribution of the sample proportions as the sample size gets bigger.<sup>5</sup> Samples of size 1 or 2 just aren't going to work very well. But the distributions of proportions of many larger samples do have histograms that are remarkably close to a Normal model.

## Assumptions and Conditions

To use a model, we usually must make some assumptions. To use the sampling distribution model for sample proportions, we need two assumptions:

**The Independence Assumption:** The sampled values must be independent of each other.

**The Sample Size Assumption:** The sample size,  $n$ , must be large enough.

Of course, assumptions are hard—often impossible—to check. That's why we *assume* them. But, as we saw in Chapter 8, we should check to see whether the assumptions are reasonable. To think about the Independence Assumption, we often wonder whether there is any reason to think that the data values might affect each other. Fortunately, we can often check *conditions* that provide information about the assumptions. Check these conditions before using the Normal to model the distribution of sample proportions:

**Randomization Condition:** If your data come from an experiment, subjects should have been randomly assigned to treatments. If you have a survey, your sample should be a simple random sample of the population. If some other sampling design was used, be sure the sampling method was not biased and that the data are representative of the population.

**10% Condition:** The sample size,  $n$ , must be no larger than 10% of the population. For national polls, the total population is usually very large, so the sample is a small fraction of the population.

**Success/Failure Condition:** The sample size has to be big enough so that we expect at least 10 successes and at least 10 failures. When  $np$  and  $nq$  are at least 10, we have enough data for sound conclusions. For the CBS survey, a "success" might be believing in ghosts. With  $p = 0.45$ , we expect  $808 \times 0.45 = 364$  successes and  $808 \times 0.55 = 444$  failures. Both are at least 10, so we certainly expect enough successes and enough failures for the condition to be satisfied.

The terms "success" and "failure" for the outcomes that have probability  $p$  and  $q$  are common in Statistics. But they are completely arbitrary labels. When we say that a disease occurs with probability  $p$ , we certainly don't mean that getting sick is a "success" in the ordinary sense of the word.

<sup>5</sup> Formally, we say the claim is true in the limit as  $n$  grows.

These last two conditions seem to conflict with each other. The **Success/Failure Condition** wants sufficient data. How much depends on  $p$ . If  $p$  is near 0.5, we need a sample of only 20 or so. If  $p$  is only 0.01, however, we'd need 1000. But the **10% Condition** says that a sample should be no larger than 10% of the population. If you're thinking, "Wouldn't a larger sample be better?" you're right of course. It's just that if the sample were more than 10% of the population, we'd need to use different methods to analyze the data. Fortunately, this isn't usually a problem in practice. Often, as in polls that sample from all U.S. adults or industrial samples from a day's production, the populations are much larger than 10 times the sample size.

## A Sampling Distribution Model for a Proportion

We've simulated repeated samples and looked at a histogram of the sample proportions. We modeled that histogram with a Normal model. Why do we bother to model it? Because this model will give us insight into how much the sample proportion can vary from sample to sample. We've simulated many of the other random samples we might have gotten. The model is an attempt to show the distribution from *all* the random samples. But how do we know that a Normal model will really work? Is this just an observation based on some simulations that *might* be approximately true some of the time?

It turns out that this model can be justified theoretically and that the larger the sample size, the better the model works. That's the result Laplace proved. We won't bother you with the math because, in this instance, it really wouldn't help your understanding.<sup>6</sup> Nevertheless, the fact that we can think of the sample proportion as a random variable taking on a different value in each random sample, and then say something this specific about the distribution of those values, is a fundamental insight—one that we will use in each of the next four chapters.

We have changed our point of view in a very important way. No longer is a proportion something we just compute for a set of data. We now see it as a random variable quantity that has a probability distribution, and thanks to Laplace we have a model for that distribution. We call that the **sampling distribution model** for the proportion, and we'll make good use of it.

**A S** *Simulation: Simulate the Sampling Distribution Model of a Proportion.* You probably don't want to work through the formal mathematical proof; a simulation is far more convincing!

We have now answered the question raised at the start of the chapter. To know how variable a sample proportion is, we need to know the proportion and the size of the sample. That's all.

### THE SAMPLING DISTRIBUTION MODEL FOR A PROPORTION

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of  $\hat{p}$  is modeled by a Normal model

with mean  $\mu(\hat{p}) = p$  and standard deviation  $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ .

Without the sampling distribution model, the rest of Statistics just wouldn't exist. Sampling models are what makes Statistics work. They inform us about the amount of variation we should expect when we sample. Suppose we spin a coin 100 times in order to decide whether it's fair or not. If we get 52 heads, we're probably not surprised. Although we'd expect 50 heads, 52 doesn't seem particularly unusual for a fair coin. But we would be surprised to see 90 heads; that might really make us doubt that the coin is fair. How about 64 heads? Harder to say. That's a case where we need the sampling distribution model. The sampling model quantifies the variability, telling us how surprising any sample proportion is. And

<sup>6</sup> The proof is pretty technical. We're not sure it helps *our* understanding all that much either.

it enables us to make informed decisions about how precise our estimate of the true proportion might be. That's exactly what we'll be doing for the rest of this book.

Sampling distribution models act as a bridge from the real world of data to the imaginary model of the statistic and enable us to say something about the population when all we have is data from the real world. This is the huge leap of Statistics. Rather than thinking about the sample proportion as a fixed quantity calculated from our data, we now think of it as a random variable—our value is just one of many we might have seen had we chosen a different random sample. By imagining what *might* happen if we were to draw many, many samples from the same population, we can learn a lot about how close the statistics computed from our one particular sample may be to the corresponding population parameters they estimate. That's the path to the *margin of error* you hear about in polls and surveys. We'll see how to determine that in the next chapter.

### FOR EXAMPLE

#### Using the sampling distribution model for proportions

The Centers for Disease Control and Prevention report that 22% of 18-year-old women in the United States have a body mass index (BMI)<sup>7</sup> of 25 or more—a value considered by the National Heart Lung and Blood Institute to be associated with increased health risk.

As part of a routine health check at a large college, the physical education department usually requires students to come in to be measured and weighed. This year, the department decided to try out a self-report system. It asked 200 randomly selected female students to report their heights and weights (from which their BMIs could be calculated). Only 31 of these students had BMIs greater than 25.

**Question:** Is this proportion of high-BMI students unusually small?

First, check the conditions:

- ✓ **Randomization Condition:** The department drew a random sample, so the respondents should be independent and randomly selected from the population.
- ✓ **10% Condition:** 200 respondents is less than 10% of all the female students at a “large college.”
- ✓ **Success/Failure Condition:** The department expected  $np = 200(0.22) = 44$  “successes” and  $nq = 200(0.78) = 156$  “failures,” both at least 10.

It's okay to use a Normal model to describe the sampling distribution of the proportion of respondents with BMIs above 25.

The phys ed department observed  $\hat{p} = \frac{31}{200} = 0.155$ .

The department expected  $E(\hat{p}) = p = 0.22$ , with  $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.22)(0.78)}{200}} = 0.029$ ,

so  $z = \frac{\hat{p} - p}{SD(\hat{p})} = \frac{0.155 - 0.22}{0.029} = -2.24$ .

By the 68–95–99.7 Rule, I know that values more than 2 standard deviations below the mean of a Normal model show up less than 2.5% of the time. Perhaps women at this college differ from the general population, or self-reporting may not provide accurate heights and weights.

<sup>7</sup> BMI = weight in kg / (height in m)<sup>2</sup>.





## JUST CHECKING

1. You want to poll a random sample of 100 students on campus to see if they are in favor of the proposed location for the new student center. Of course, you'll get just one number, your sample proportion,  $\hat{p}$ . But if you imagined all the possible samples of 100 students you could draw and imagined the histogram of all the sample proportions from these samples, what shape would it have?
2. Where would the center of that histogram be?
3. If you think that about half the students are in favor of the plan, what would the standard deviation of the sample proportions be?

### STEP-BY-STEP EXAMPLE

### Working with Sampling Distribution Models for Proportions

Suppose that about 13% of the population is left-handed.<sup>8</sup> A 200-seat school auditorium has been built with 15 “lefty seats,” seats that have the built-in desk on the left rather than the right arm of the chair. (For the right-handed readers among you, have you ever tried to take notes in a chair with the desk on the left side?)

**Question:** In a class of 90 students, what's the probability that there will not be enough seats for the left-handed students?



**Plan** State what we want to know.

**Model** Think about the assumptions and check the conditions.

You might be able to think of cases where the **Independence Assumption** is not plausible—for example, if the students are all related, or if they were selected for being left- or right-handed. But for a random sample, the assumption of independence seems reasonable.

I want to find the probability that in a group of 90 students, more than 15 will be left-handed. Since 15 out of 90 is 16.7%, I need the probability of finding more than 16.7% left-handed students out of a sample of 90 if the proportion of lefties is 13%.

- ✓ **Independence Assumption:** It is reasonable to assume that the probability that one student is left-handed is not changed by the fact that another student is right- or left-handed.
- ✓ **Randomization Condition:** The 90 students in the class can be thought of as a random sample of students.
- ✓ **10% Condition:** 90 is surely less than 10% of the population of all students. (Even if the school itself is small, I'm thinking of the population of all possible students who could have gone to the school.)
- ✓ **Success/Failure Condition:**

$$np = 90(0.13) = 11.7 \geq 10$$

$$nq = 90(0.87) = 78.3 \geq 10$$

<sup>8</sup> Actually, it's quite difficult to get an accurate estimate of the proportion of lefties in the population. Estimates range from 8% to 15%.

State the parameters and the sampling distribution model.

The population proportion is  $p = 0.13$ . The conditions are satisfied, so I'll model the sampling distribution of  $\hat{p}$  with a Normal model with mean 0.13 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{90}} \approx 0.035$$

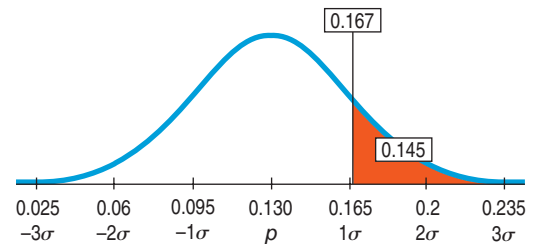
My model for  $\hat{p}$  is  $N(0.13, 0.035)$ .

**SHOW**

**Plot** Make a picture. Sketch the model and shade the area we're interested in, in this case the area to the right of 0.167.

**Mechanics** Use the standard deviation as a ruler to find the z-score of the cutoff proportion. We see that 16.7% lefties would be just over one standard deviation above the mean.

Find the resulting probability from a table of Normal probabilities, a computer program, or a calculator.



$$z = \frac{\hat{p} - p}{SD(\hat{p})} = \frac{0.167 - 0.13}{0.035} = 1.06$$

$$P(\hat{p} > 0.167) = P(z > 1.06) = 0.1446$$

**TELL**

**Conclusion** Interpret the probability in the context of the question.

There is about a 14.5% chance that there will not be enough seats for the left-handed students in the class.

## What About Quantitative Data?

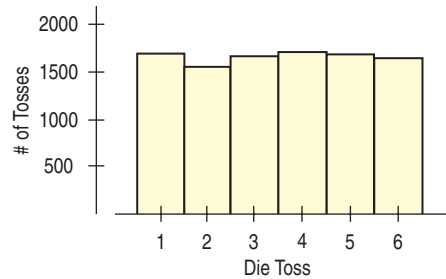
Proportions summarize categorical variables. And the Normal sampling distribution model looks like it is going to be very useful. But can we do something similar with quantitative data?

Of course we can (or we wouldn't have asked). Even more remarkable, not only can we use all of the same concepts, but almost the same model, too.

What are the concepts? We know that when we sample at random or randomize an experiment, the results we get will vary from sample-to-sample and from experiment-to-experiment. The Normal model seems an incredibly simple way to summarize all that variation. Could something that simple work for means? We won't keep you in suspense. It turns out that means also have a sampling distribution that we can model with a Normal model. And it turns out that Laplace's theoretical result applies to means, too. As we did with proportions, we can get some insight from a simulation.

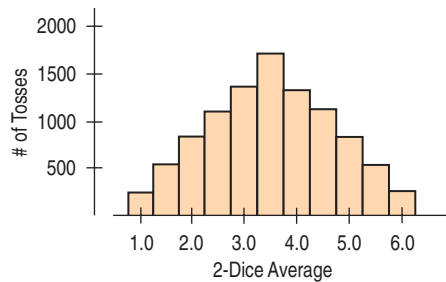
## Simulating the Sampling Distribution of a Mean

Here's a simple simulation. Let's start with one fair die. If we toss this die 10,000 times, what should the histogram of the numbers on the face of the die look like? Here are the results of a simulated 10,000 tosses:



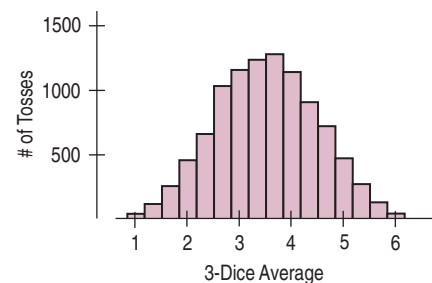
Now let's toss a *pair* of dice and record the average of the two. If we repeat this (or at least simulate repeating it) 10,000 times, recording the average of each pair, what will the histogram of these 10,000 averages look like? Before you look, think a minute. Is getting an average of 1 on *two* dice as likely as getting an average of 3 or 3.5?

Let's see:



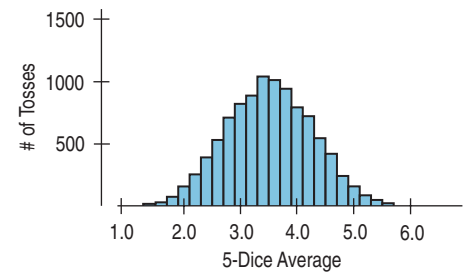
We're much more likely to get an average near 3.5 than we are to get one near 1 or 6. Without calculating those probabilities exactly, it's fairly easy to see that the *only* way to get an average of 1 is to get two 1's. To get a total of 7 (for an average of 3.5), though, there are many more possibilities. This distribution even has a name: the *triangular* distribution.

What if we average 3 dice? We'll simulate 10,000 tosses of 3 dice and take their average:



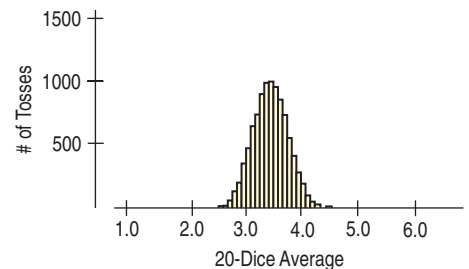
What's happening? First notice that it's getting harder to have averages near the ends. Getting an average of 1 or 6 with 3 dice requires all three to come up 1 or 6, respectively. That's less likely than for 2 dice to come up both 1 or both 6. The distribution is being pushed toward the middle. But what's happening to the shape? (This distribution doesn't have a name, as far as we know.)

Let's continue this simulation to see what happens with larger samples. Here's a histogram of the averages for 10,000 tosses of 5 dice:



The pattern is becoming clearer. Two things continue to happen. The first fact we knew already from the Law of Large Numbers. It says that as the sample size (number of dice) gets larger, each sample average is more likely to be closer to the population mean. So, we see the shape continuing to tighten around 3.5. But the shape of the distribution is the surprising part. It's becoming bell-shaped. And not just bell-shaped; it's approaching the Normal model.

Are you convinced? Let's skip ahead and try 20 dice. The histogram of averages for 10,000 throws of 20 dice looks like this:



Now we see the Normal shape again (and notice how much smaller the spread is). But can we count on this happening for situations other than dice throws? What kinds of sample means have sampling distributions that we can model with a Normal model? It turns out that Normal models work well amazingly often.

**A S** **Activity: The Sampling Distribution Model for Means.** Don't just sit there reading about the simulation—do it yourself.

## The Fundamental Theorem of Statistics

The dice simulation may look like a special situation, but it turns out that what we saw with dice is true for means of repeated samples for almost every situation. When we looked at the sampling distribution of a proportion, we had to check only a few conditions. For means, the result is even more remarkable. *There are almost no conditions at all.*

Let's say that again: The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows. All we need is for the observations to be independent and collected with randomization. We don't even care about the shape of the population distribution!<sup>9</sup> This surprising fact is the result Laplace proved in a fairly general form in 1810. At the time, Laplace's theorem caused quite a stir (at least in mathematics circles) because it is so unintuitive. Laplace's result is called the **Central Limit Theorem**<sup>10</sup> (CLT).

*"The theory of probabilities is at bottom nothing but common sense reduced to calculus."*

—Laplace, in *Théorie analytique des probabilités*, 1812

<sup>9</sup> OK, one technical condition. The data must come from a population with a finite variance. You probably can't imagine a population with an infinite variance, but statisticians can construct such things, so we have to discuss them in footnotes like this. It really makes no difference in how you think about the important stuff, so you can just forget we mentioned it.

<sup>10</sup> The word "central" in the name of the theorem means "fundamental." It doesn't refer to the center of a distribution.

Laplace was one of the greatest scientists and mathematicians of his time. In addition to his contributions to probability and statistics, he published many new results in mathematics, physics, and astronomy (where his nebular theory was one of the first to describe the formation of the solar system in much the way it is understood today). He also played a leading role in establishing the metric system of measurement.

His brilliance, though, sometimes got him into trouble. A visitor to the Académie des Sciences in Paris reported that Laplace let it be widely known that he considered himself the best mathematician in France. The effect of this on his colleagues was not eased by the fact that Laplace was right.

#### TI-*n*spire

**The Central Limit Theorem.** See the sampling distribution of sample means take shape as you choose sample after sample.

Why should the Normal model show up again for the sampling distribution of means as well as proportions? We're not going to try to persuade you that it is obvious, clear, simple, or straightforward. In fact, the CLT is surprising and a bit weird. Not only does the distribution of means of many random samples get closer and closer to a Normal model as the sample size grows, *this is true regardless of the shape of the population distribution!* Even if we sample from a skewed or bimodal population, the Central Limit Theorem tells us that means of repeated random samples will tend to follow a Normal model as the sample size grows. Of course, you won't be surprised to learn that it works better and faster the closer the population distribution is to a Normal model. And it works better for larger samples. If the data come from a population that's exactly Normal to start with, then the observations themselves are Normal. If we take samples of size 1, their "means" are just the observations—so, of course, they have Normal sampling distribution. But now suppose the population distribution is very skewed (like the CEO data from Chapter 5, for example). The CLT works, although it may take a sample size of dozens or even hundreds of observations for the Normal model to work well.

For example, think about a really bimodal population, one that consists of only 0's and 1's. The CLT says that even means of samples from this population will follow a Normal sampling distribution model. But wait. Suppose we have a categorical variable and we assign a 1 to each individual in the category and a 0 to each individual not in the category. And then we find the mean of these 0's and 1's. That's the same as counting the number of individuals who are in the category and dividing by  $n$ . That mean will be . . . the *sample proportion*,  $\hat{p}$ , of individuals who are in the category (a "success"). So maybe it wasn't so surprising after all that proportions, like means, have Normal sampling distribution models; they are actually just a special case of Laplace's remarkable theorem. Of course, for such an extremely bimodal population, we'll need a reasonably large sample size—and that's where the special conditions for proportions come in.

#### THE CENTRAL LIMIT THEOREM (CLT)

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.

## Assumptions and Conditions

### AS

**Activity: The Central Limit Theorem.** Does it really work for samples from non-Normal populations?

The CLT requires essentially the same assumptions as we saw for modelling proportions:

**Independence Assumption:** The sampled values must be independent of each other.

**Sample Size Assumption:** The sample size must be sufficiently large.

We can't check these directly, but we can think about whether the **Independence Assumption** is plausible. We can also check some related conditions:

**Randomization Condition:** The data values must be sampled randomly, or the concept of a sampling distribution makes no sense.

**10% Condition:** When the sample is drawn without replacement (as is usually the case), the sample size,  $n$ , should be no more than 10% of the population.

**Large Enough Sample Condition:** Although the CLT tells us that a Normal model is useful in thinking about the behavior of sample means when the

sample size is large enough, it doesn't tell us how large a sample we need. The truth is, it depends; there's no one-size-fits-all rule. If the population is unimodal and symmetric, even a fairly small sample is okay. If the population is strongly skewed, like the compensation for CEOs we looked at in Chapter 5, it can take a pretty large sample to allow use of a Normal model to describe the distribution of sample means. For now you'll just need to think about your sample size in the context of what you know about the population, and then tell whether you believe the **Large Enough Sample Condition** has been met.

## But Which Normal?

**A S**

**Activity: The Standard Deviation of Means.** Experiment to see how the variability of the mean changes with the sample size.

The CLT says that the sampling distribution of any mean or proportion is approximately Normal. But which Normal model? We know that any Normal is specified by its mean and standard deviation. For proportions, the sampling distribution is centered at the population proportion. For means, it's centered at the population mean. What else would we expect?

What about the standard deviations, though? We noticed in our dice simulation that the histograms got narrower as we averaged more and more dice together. This shouldn't be surprising. Means vary less than the individual observations. Think about it for a minute. Which would be more surprising, having *one* person in your Statistics class who is over 6'9" tall or having the *mean* of 100 students taking the course be over 6'9"? The first event is fairly rare.<sup>11</sup> You may have seen somebody this tall in one of your classes sometime. But finding a class of 100 whose mean height is over 6'9" tall just won't happen. Why? Because *means have smaller standard deviations than individuals*.

How much smaller? Well, we have good news and bad news. The good news is that the standard deviation of  $\bar{y}$  falls as the sample size grows. The bad news is that it doesn't drop as fast as we might like. It only goes down by the *square root* of the sample size. Why? The Math Box will show you that the Normal model for the sampling distribution of the mean has a standard deviation equal to

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the standard deviation of the population. To emphasize that this is a standard deviation *parameter* of the sampling distribution model for the sample mean,  $\bar{y}$ , we write  $SD(\bar{y})$  or  $\sigma(\bar{y})$ .

**A S**

**Activity: The Sampling Distribution of the Mean.** The CLT tells us what to expect. In this activity you can work with the CLT or simulate it if you prefer.

### THE SAMPLING DISTRIBUTION MODEL FOR A MEAN (CLT)

When a random sample is drawn from any population with mean  $\mu$  and standard deviation  $\sigma$ , its sample mean,  $\bar{y}$ , has a sampling distribution

with the same *mean*  $\mu$  but whose *standard deviation* is  $\frac{\sigma}{\sqrt{n}}$  (and we write

$\sigma(\bar{y}) = SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ ). No matter what population the random sample comes

from, the *shape* of the sampling distribution is approximately Normal as long as the sample size is large enough. The larger the sample used, the more closely the Normal approximates the sampling distribution for the mean.

<sup>11</sup> If students are a random sample of adults, fewer than 1 out of 10,000 should be taller than 6'9". Why might college students not really be a random sample with respect to height? Even if they're not a perfectly random sample, a college student over 6'9" tall is still rare.

## MATH BOX

We know that  $\bar{y}$  is a sum divided by  $n$ :

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \cdots + y_n}{n}.$$

As we saw in Chapter 16, when a random variable is divided by a constant its variance is divided by the *square* of the constant:

$$\text{Var}(\bar{y}) = \frac{\text{Var}(y_1 + y_2 + y_3 + \cdots + y_n)}{n^2}.$$

To get our sample, we draw the  $y$ 's randomly, ensuring they are independent. For independent random variables, variances add:

$$\text{Var}(\bar{y}) = \frac{\text{Var}(y_1) + \text{Var}(y_2) + \text{Var}(y_3) + \cdots + \text{Var}(y_n)}{n^2}.$$

All  $n$  of the  $y$ 's were drawn from our population, so they all have the same variance,  $\sigma^2$ :

$$\text{Var}(\bar{y}) = \frac{\sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The standard deviation of  $\bar{y}$  is the square root of this variance:

$$SD(\bar{y}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

We now have two closely related sampling distribution models that we can use when the appropriate assumptions and conditions are met. Which one we use depends on which kind of data we have:

- ▶ When we have categorical data, we calculate a sample proportion,  $\hat{p}$ ; the sampling distribution of this random variable has a Normal model with a mean at the true proportion (“Greek letter”)  $p$  and a standard deviation of  $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}}$ . We’ll use this model in Chapters 19 through 22.
- ▶ When we have quantitative data, we calculate a sample mean,  $\bar{y}$ ; the sampling distribution of this random variable has a Normal model with a mean at the true mean,  $\mu$ , and a standard deviation of  $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ . We’ll use this model in Chapters 23, 24, and 25.

The means of these models are easy to remember, so all you need to be careful about is the standard deviations. Remember that these are standard deviations of the *statistics*  $\hat{p}$  and  $\bar{y}$ . They both have a square root of  $n$  in the denominator. That tells us that the larger the sample, the less either statistic will vary. The only difference is in the numerator. If you just start by writing  $SD(\bar{y})$  for quantitative data and  $SD(\hat{p})$  for categorical data, you’ll be able to remember which formula to use.

## FOR EXAMPLE

## Using the CLT for means

**Recap:** A college physical education department asked a random sample of 200 female students to self-report their heights and weights, but the percentage of students with body mass indexes over 25 seemed suspiciously low. One possible explanation may be that the respondents “shaded” their weights down a bit. The CDC reports that the mean weight of 18-year-old women is 143.74 lb, with a standard deviation of 51.54 lb, but these 200 randomly selected women reported a mean weight of only 140 lb.

**Question:** Based on the Central Limit Theorem and the 68–95–99.7 Rule, does the mean weight in this sample seem exceptionally low, or might this just be random sample-to-sample variation?

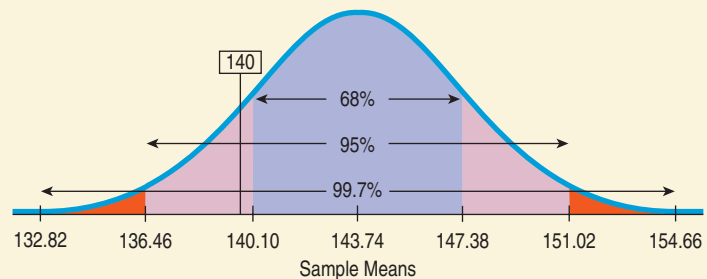
The conditions check out okay:

- ✓ **Randomization Condition:** The women were a random sample and their weights can be assumed to be independent.
- ✓ **10% Condition:** They sampled fewer than 10% of all women at the college.
- ✓ **Large Enough Sample Condition:** The distribution of college women’s weights is likely to be unimodal and reasonably symmetric, so the CLT applies to means of even small samples; 200 values is plenty.

The sampling model for sample means is approximately Normal with  $E(\bar{y}) = 143.7$  and

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{51.54}{\sqrt{200}} = 3.64. \text{ The expected}$$

distribution of sample means is:



The 68–95–99.7 Rule suggests that although the reported mean weight of 140 pounds is somewhat lower than expected, it does not appear to be unusual. Such variability is not all that extraordinary for samples of this size.

## STEP-BY-STEP EXAMPLE

## Working with the Sampling Distribution Model for the Mean

The Centers for Disease Control and Prevention reports that the mean weight of adult men in the United States is 190 lb with a standard deviation of 59 lb.<sup>12</sup>

**Question:** An elevator in our building has a weight limit of 10 persons or 2500 lb. What’s the probability that if 10 men get on the elevator, they will overload its weight limit?

THINK

**Plan** State what we want to know.

Asking the probability that the total weight of a sample of 10 men exceeds 2500 pounds is equivalent to asking the probability that their mean weight is greater than 250 pounds.

<sup>12</sup> Cynthia L. Ogden, Cheryl D. Fryar, Margaret D. Carroll, and Katherine M. Flegal, *Mean Body Weight, Height, and Body Mass Index, United States 1960–2002, Advance Data from Vital and Health Statistics Number 347*, Oct. 27, 2004. <https://www.cdc.gov/nchs>



**Model** Think about the assumptions and check the conditions.

Note that if the sample were larger we'd be less concerned about the shape of the distribution of all weights.

State the parameters and the sampling model.

- ✓ **Independence Assumption:** It's reasonable to think that the weights of 10 randomly sampled men will be independent of each other. (But there could be exceptions—for example, if they were all from the same family or if the elevator were in a building with a diet clinic!)
- ✓ **Randomization Condition:** I'll assume that the 10 men getting on the elevator are a random sample from the population.
- ✓ **10% Condition:** 10 men is surely less than 10% of the population of possible elevator riders.
- ✓ **Large Enough Sample Condition:** I suspect the distribution of population weights is roughly unimodal and symmetric, so my sample of 10 men seems large enough.

The mean for all weights is  $\mu = 190$  and the standard deviation is  $\sigma = 59$  pounds. Since the conditions are satisfied, the CLT says that the sampling distribution of  $\bar{y}$  has a Normal model with mean 190 and standard deviation

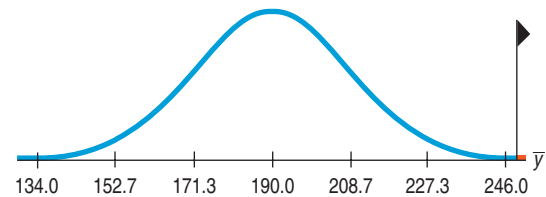
$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{59}{\sqrt{10}} \approx 18.66$$

SHOW

**Plot** Make a picture. Sketch the model and shade the area we're interested in. Here the mean weight of 250 pounds appears to be far out on the right tail of the curve.

**Mechanics** Use the standard deviation as a ruler to find the z-score of the cutoff mean weight. We see that an average of 250 pounds is more than 3 standard deviations above the mean.

Find the resulting probability from a table of Normal probabilities such as Table Z, a computer program, or a calculator.



$$z = \frac{\bar{y} - \mu}{SD(\bar{y})} = \frac{250 - 190}{18.66} = 3.21$$

$$P(\bar{y} > 250) = P(z > 3.21) = 0.0007$$

TELL

**Conclusion** Interpret your result in the proper context, being careful to relate it to the original question.

The chance that a random collection of 10 men will exceed the elevator's weight limit is only 0.0007. So, if they are a random sample, it is quite unlikely that 10 people will exceed the total weight allowed on the elevator.

## About Variation

“The  $n$ 's justify the means.”

—Apocryphal  
statistical saying

Means vary less than individual data values. That makes sense. If the same test is given to many sections of a large course and the class average is, say, 80%, some students may score 95% because individual scores vary a lot. But we'd be shocked (and pleased!) if the *average* score of the students in any section was 95%. Averages are much less variable. Not only do group averages vary less than individual values, but common sense suggests that averages should be more consistent for larger groups. The Central Limit Theorem confirms this hunch; the fact that  $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$  has  $n$  in the denominator shows that the variability of sample means decreases as the sample size increases. There's a catch, though. The standard deviation of the sampling distribution declines only with the square root of the sample size and not, for example, with  $1/n$ .

The mean of a random sample of 4 has half  $\left(\frac{1}{\sqrt{4}} = \frac{1}{2}\right)$  the standard deviation of an individual data value. To cut the standard deviation in half again, we'd need a sample of 16, and a sample of 64 to halve it once more.

If only we had a much larger sample, we could get the standard deviation of the sampling distribution *really* under control so that the sample mean could tell us still more about the unknown population mean, but larger samples cost more and take longer to survey. And while we're gathering all that extra data, the population itself may change, or a news story may alter opinions. There are practical limits to most sample sizes. As we shall see, that nasty square root limits how much we can make a sample tell about the population. This is an example of something that's known as the Law of Diminishing Returns.

**A Billion Dollar Misunderstanding?** In the late 1990s the Bill and Melinda Gates Foundation began funding an effort to encourage the breakup of large schools into smaller schools. Why? It had been noticed that smaller schools were more common among the best-performing schools than one would expect. In time, the Annenberg Foundation, the Carnegie Corporation, the Center for Collaborative Education, the Center for School Change, Harvard's Change Leadership Group, the Open Society Institute, Pew Charitable Trusts, and the U.S. Department of Education's Smaller Learning Communities Program all supported the effort. Well over a billion dollars was spent to make schools smaller.

But was it all based on a misunderstanding of sampling distributions? Statisticians Howard Wainer and Harris Zwerling<sup>13</sup> looked at the mean test scores of schools in Pennsylvania. They found that indeed 12% of the top-scoring 50 schools were from the smallest 3% of Pennsylvania schools—substantially more than the 3% we'd naively expect. But then they looked at the *bottom* 50. There they found that 18% were small schools! The explanation? Mean test scores are, well, means. We are looking at a rough real-world simulation in which each school is a trial. Even if all Pennsylvania schools were equivalent, we'd expect their mean scores to vary. How much? The CLT tells us that means of test scores vary according to  $\frac{\sigma}{\sqrt{n}}$ . Smaller schools have (by definition) smaller  $n$ 's, so the sampling distributions of their mean scores naturally have larger standard deviations. It's natural, then, that small schools have both higher and lower mean scores.

<sup>13</sup> Wainer, H. and Zwerling, H., “Legal and empirical evidence that smaller schools do not improve student achievement,” *The Phi Delta Kappan* 2006 87:300–303. Discussed in Howard Wainer, “The Most Dangerous Equation,” *American Scientist*, May–June 2007, pp. 249–256; also at [www.Americanscientist.org](http://www.Americanscientist.org).

On October 26, 2005, *The Seattle Times* reported:

*[T]he Gates Foundation announced last week it is moving away from its emphasis on converting large high schools into smaller ones and instead giving grants to specially selected school districts with a track record of academic improvement and effective leadership. Education leaders at the Foundation said they concluded that improving classroom instruction and mobilizing the resources of an entire district were more important first steps to improving high schools than breaking down the size.*

## The Real World and the Model World

Be careful. We have been slipping smoothly between the real world, in which we draw random samples of data, and a magical mathematical model world, in which we describe how the sample means and proportions we observe in the real world behave as random variables in all the random samples that we might have drawn. Now we have *two* distributions to deal with. The first is the real-world distribution of the sample, which we might display with a histogram (for quantitative data) or with a bar chart or table (for categorical data). The second is the math world *sampling distribution model* of the statistic, a Normal model based on the Central Limit Theorem. Don't confuse the two.

For example, don't mistakenly think the CLT says that the *data* are Normally distributed as long as the sample is large enough. In fact, as samples get larger, we expect the distribution of the data to look more and more like the population from which they are drawn—skewed, bimodal, whatever—but not necessarily Normal. You can collect a sample of CEO salaries for the next 1000 years,<sup>14</sup> but the histogram will never look Normal. It will be skewed to the right. The Central Limit Theorem doesn't talk about the distribution of the data from the sample. It talks about the sample *means* and sample *proportions* of many different random samples drawn from the same population. Of course, the CLT does require that the sample be big enough when the population shape is not unimodal and symmetric, but the fact that, even then, a Normal model is useful is still a very surprising and powerful result.



### JUST CHECKING

4. Human gestation times have a mean of about 266 days, with a standard deviation of about 16 days. If we record the gestation times of a sample of 100 women, do we know that a histogram of the times will be well modeled by a Normal model?
5. Suppose we look at the *average* gestation times for a sample of 100 women. If we imagined all the possible random samples of 100 women we could take and looked at the histogram of all the sample means, what shape would it have?
6. Where would the center of that histogram be?
7. What would be the standard deviation of that histogram?

<sup>14</sup> Don't forget to adjust for inflation.

## Sampling Distribution Models

Let's summarize what we've learned about sampling distributions. At the heart is the idea that *the statistic itself is a random variable*. We can't know what our statistic will be because it comes from a random sample. It's just one instance of something that happened for our particular random sample. A different random sample would have given a different result. This sample-to-sample variability is what generates the sampling distribution. The sampling distribution shows us the distribution of possible values that the statistic could have had.

We could simulate that distribution by pretending to take lots of samples. Fortunately, for the mean and the proportion, the CLT tells us that we can model their sampling distribution directly with a Normal model.

The two basic truths about sampling distributions are:

1. Sampling distributions arise because samples vary. Each random sample will contain different cases and, so, a different value of the statistic.
2. Although we can always simulate a sampling distribution, the Central Limit Theorem saves us the trouble for means and proportions.

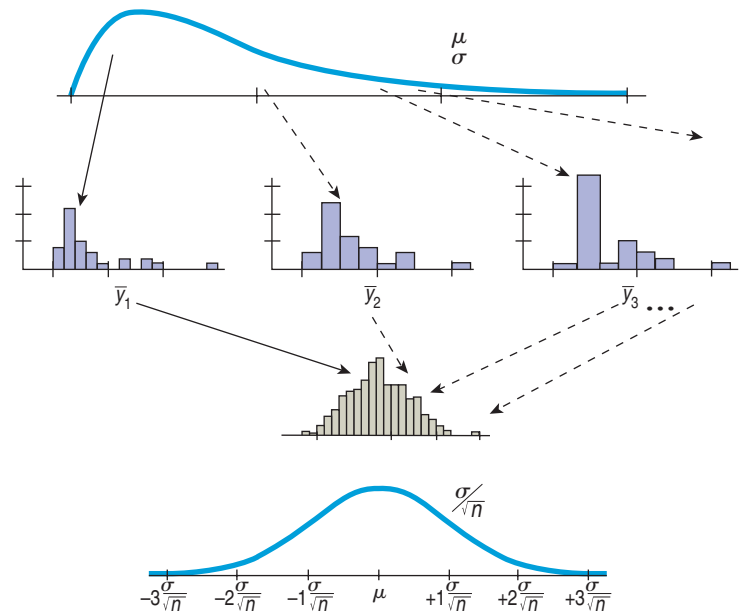
Here's a picture showing the process going into the sampling distribution model:

**A S** **Simulation: The CLT for Real Data.** Why settle for a picture when you can see it in action?

**FIGURE 18.5**

We start with a population model, which can have any shape. It can even be bimodal or skewed (as this one is). We label the mean of this model  $\mu$  and its standard deviation,  $\sigma$ .

We draw one real sample (solid line) of size  $n$  and show its histogram and summary statistics. We imagine (or simulate) drawing many other samples (dotted lines), which have their own histograms and summary statistics.



We (imagine) gathering all the means into a histogram.

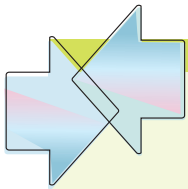
The CLT tells us we can model the shape of this histogram with a Normal model. The mean of this Normal is  $\mu$ , and the standard deviation is  $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ .

### WHAT CAN GO WRONG?

- ▶ **Don't confuse the sampling distribution with the distribution of the sample.** When you take a sample, you always look at the distribution of the values, usually with a histogram, and you may calculate summary statistics. Examining the distribution of the sample data is wise. But that's not the sampling distribution. The sampling distribution is an imaginary collection of all the values that a statistic *might* have taken for all possible random samples—the one you got and the ones that you didn't get. We use the sampling distribution model to make statements about how the statistic varies.

(continued)

- ▶ **Beware of observations that are not independent.** The CLT depends crucially on the assumption of independence. If our elevator riders are related, are all from the same school (for example, an elementary school), or in some other way aren't a random sample, then the statements we try to make about the mean are going to be wrong. Unfortunately, this isn't something you can check in your data. You have to think about how the data were gathered. Good sampling practice and well-designed randomized experiments ensure independence.
- ▶ **Watch out for small samples from skewed populations.** The CLT assures us that the sampling distribution model is Normal if  $n$  is large enough. If the population is nearly Normal, even small samples (like our 10 elevator riders) work. If the population is very skewed, then  $n$  will have to be large before the Normal model will work well. If we sampled 15 or even 20 CEOs and used  $\bar{y}$  to make a statement about the mean of all CEOs' compensation, we'd likely get into trouble because the underlying data distribution is so skewed. Unfortunately, there's no good rule of thumb.<sup>15</sup> It just depends on how skewed the data distribution is. Always plot the data to check.



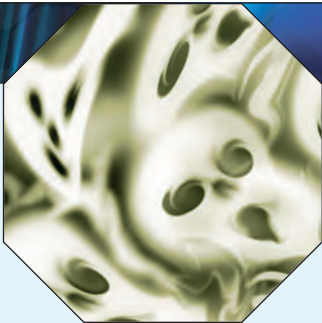
## CONNECTIONS

The concept of a sampling distribution connects to almost everything we have done. The fundamental connection is to the deliberate application of randomness in random sampling and randomized comparative experiments. If we didn't employ randomness to generate unbiased data, then repeating the data collection would just get the same data values again (with perhaps a few new measurement or recording errors). The distribution of statistic values arises directly because different random samples and randomized experiments would generate different statistic values.

The connection to the Normal distribution is obvious. We first introduced the Normal model before because it was "nice." As a unimodal, symmetric distribution with 99.7% of its area within three standard deviations of the mean, the Normal model is easy to work with. Now we see that the Normal holds a special place among distributions because we can use it to model the sampling distributions of the mean and the proportion.

We use simulation to understand sampling distributions. In fact, some important sampling distributions were discovered first by simulation.

## WHAT HAVE WE LEARNED?



Way back in Chapter 1 we said that Statistics is about variation. We know that no sample fully and exactly describes the population; sample proportions and means will vary from sample to sample. That's sampling error (or, better, sampling variability). We know it will always be present—indeed, the world would be a boring place if variability didn't exist. You might think that sampling variability would prevent us from learning anything reliable about a population by looking at a sample, but that's just not so. The fortunate fact is that sampling variability is not just unavoidable—it's predictable!

<sup>15</sup> For proportions, of course, there is a rule: the **Success/Failure Condition**. That works for proportions because the standard deviation of a proportion is linked to its mean.

We've learned how the Central Limit Theorem describes the behavior of sample proportions—shape, center, and spread—as long as certain assumptions and conditions are met. The sample must be independent, random, and large enough that we expect at least 10 successes and failures. Then:

- ▶ The sampling distribution (the imagined histogram of the proportions from all possible samples) is shaped like a Normal model.
- ▶ The mean of the sampling model is the true proportion in the population.
- ▶ The standard deviation of the sample proportions is  $\sqrt{\frac{pq}{n}}$ .

And we've learned to describe the behavior of sample means as well, based on this amazing result known as the Central Limit Theorem—the Fundamental Theorem of Statistics. Again the sample must be independent and random—no surprise there—and needs to be larger if our data come from a population that's not roughly unimodal and symmetric. Then:

- ▶ Regardless of the shape of the original population, the shape of the distribution of the means of all possible samples can be described by a Normal model, provided the samples are large enough.
- ▶ The center of the sampling model will be the true mean of the population from which we took the sample.
- ▶ The standard deviation of the sample means is the population's standard deviation divided by the square root of the sample size,  $\frac{\sigma}{\sqrt{n}}$ .

## Terms

Sampling distribution model	413. Different random samples give different values for a statistic. The sampling distribution model shows the behavior of the statistic over all the possible samples for the same size $n$ .
Sampling variability Sampling error	414. The variability we expect to see from one random sample to another. It is sometimes called sampling error, but sampling variability is the better term.
Sampling distribution model for a proportion	416. If assumptions of independence and random sampling are met, and we expect at least 10 successes and 10 failures, then the sampling distribution of a proportion is modeled by a Normal model with a mean equal to the true proportion value, $p$ , and a standard deviation equal to $\sqrt{\frac{pq}{n}}$ .
Central Limit Theorem	421. The Central Limit Theorem (CLT) states that the sampling distribution model of the sample mean (and proportion) from a random sample is approximately Normal for large $n$ , <i>regardless of the distribution of the population, as long as the observations are independent.</i>
Sampling distribution model for a mean	423. If assumptions of independence and random sampling are met, and the sample size is large enough, the sampling distribution of the sample mean is modeled by a Normal model with a mean equal to the population mean, $\mu$ , and a standard deviation equal to $\frac{\sigma}{\sqrt{n}}$ .

## Skills

THINK

- ▶ Understand that the variability of a statistic (as measured by the standard deviation of its sampling distribution) depends on the size of the sample. Statistics based on larger samples are less variable.

SHOW

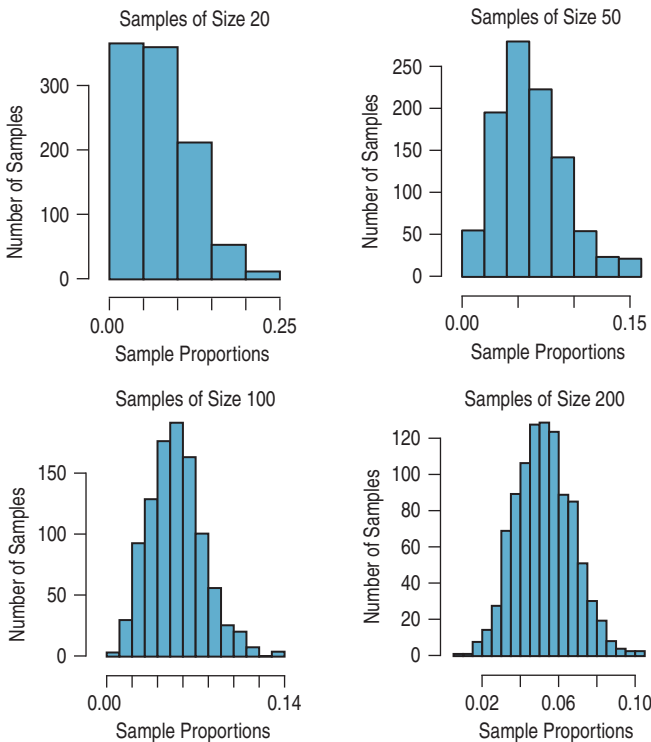
- ▶ Understand that the Central Limit Theorem gives the sampling distribution model of the mean for sufficiently large samples regardless of the underlying population.
- ▶ Be able to demonstrate a sampling distribution by simulation.
- ▶ Be able to use a sampling distribution model to make simple statements about the distribution of a proportion or mean under repeated sampling.

TELL

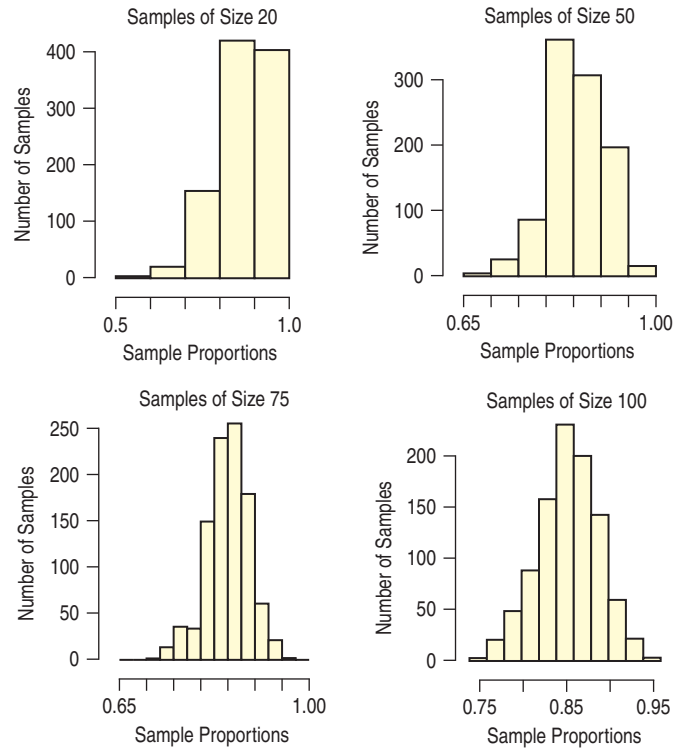
- ▶ Be able to interpret a sampling distribution model as describing the values taken by a statistic in all possible realizations of a sample or randomized experiment under the same conditions.

## EXERCISES

1. **Send money.** When they send out their fundraising letter, a philanthropic organization typically gets a return from about 5% of the people on their mailing list. To see what the response rate might be for future appeals, they did a simulation using samples of size 20, 50, 100, and 200. For each sample size, they simulated 1000 mailings with success rate  $p = 0.05$  and constructed the histogram of the 1000 sample proportions, shown below. Explain how these histograms demonstrate what the Central Limit Theorem says about the sampling distribution model for sample proportions. Be sure to talk about shape, center, and spread.



2. **Character recognition.** An automatic character recognition device can successfully read about 85% of handwritten credit card applications. To estimate what might happen when this device reads a stack of applications, the company did a simulation using samples of size 20, 50, 75, and 100. For each sample size, they simulated 1000 samples with success rate  $p = 0.85$  and constructed the histogram of the 1000 sample proportions, shown here. Explain how these histograms demonstrate what the Central Limit Theorem says about the sampling distribution model for sample proportions. Be sure to talk about shape, center, and spread.



3. **Send money, again.** The philanthropic organization in Exercise 1 expects about a 5% success rate when they send fundraising letters to the people on their mailing list. In Exercise 1 you looked at the histograms showing distributions of sample proportions from 1000 simulated mailings for samples of size 20, 50, 100, and 200. The sample statistics from each simulation were as follows:

$n$	mean	st. dev.
20	0.0497	0.0479
50	0.0516	0.0309
100	0.0497	0.0215
200	0.0501	0.0152

- According to the Central Limit Theorem, what should the theoretical mean and standard deviations be for these sample sizes?
- How close are those theoretical values to what was observed in these simulations?
- Looking at the histograms in Exercise 1, at what sample size would you be comfortable using the Normal model as an approximation for the sampling distribution?
- What does the Success/Failure Condition say about the choice you made in part c?

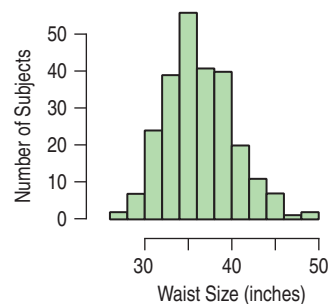
4. **Character recognition, again.** The automatic character recognition device discussed in Exercise 2 successfully reads about 85% of handwritten credit card applications. In Exercise 2 you looked at the histograms showing distributions of sample proportions from 1000 simulated samples of size 20, 50, 75, and 100. The sample statistics from each simulation were as follows:

$n$	mean	st. dev.
20	0.8481	0.0803
50	0.8507	0.0509
75	0.8481	0.0406
100	0.8488	0.0354

- a) According to the Central Limit Theorem, what should the theoretical mean and standard deviations be for these sample sizes?
- b) How close are those theoretical values to what was observed in these simulations?
- c) Looking at the histograms in Exercise 2, at what sample size would you be comfortable using the Normal model as an approximation for the sampling distribution?
- d) What does the Success/Failure Condition say about the choice you made in part c?
5. **Coin tosses.** In a large class of introductory Statistics students, the professor has each person toss a coin 16 times and calculate the proportion of his or her tosses that were heads. The students then report their results, and the professor plots a histogram of these several proportions.
- a) What shape would you expect this histogram to be? Why?
- b) Where do you expect the histogram to be centered?
- c) How much variability would you expect among these proportions?
- d) Explain why a Normal model should not be used here.
6. **M&M's.** The candy company claims that 10% of the M&M's it produces are green. Suppose that the candies are packaged at random in small bags containing about 50 M&M's. A class of elementary school students learning about percents opens several bags, counts the various colors of the candies, and calculates the proportion that are green.
- a) If we plot a histogram showing the proportions of green candies in the various bags, what shape would you expect it to have?
- b) Can that histogram be approximated by a Normal model? Explain.
- c) Where should the center of the histogram be?
- d) What should the standard deviation of the proportion be?
7. **More coins.** Suppose the class in Exercise 5 repeats the coin-tossing experiment.
- a) The students toss the coins 25 times each. Use the 68–95–99.7 Rule to describe the sampling distribution model.
- b) Confirm that you can use a Normal model here.
- c) They increase the number of tosses to 64 each. Draw and label the appropriate sampling distribution model. Check the appropriate conditions to justify your model.
- d) Explain how the sampling distribution model changes as the number of tosses increases.
8. **Bigger bag.** Suppose the class in Exercise 6 buys bigger bags of candy, with 200 M&M's each. Again the students calculate the proportion of green candies they find.
- a) Explain why it's appropriate to use a Normal model to describe the distribution of the proportion of green M&M's they might expect.
- b) Use the 68–95–99.7 Rule to describe how this proportion might vary from bag to bag.
- c) How would this model change if the bags contained even more candies?
9. **Just (un)lucky?** One of the students in the introductory Statistics class in Exercise 7 claims to have tossed her coin 200 times and found only 42% heads. What do you think of this claim? Explain.
10. **Too many green ones?** In a really large bag of M&M's, the students in Exercise 8 found 500 candies, and 12% of them were green. Is this an unusually large proportion of green M&M's? Explain.
11. **Speeding.** State police believe that 70% of the drivers traveling on a major interstate highway exceed the speed limit. They plan to set up a radar trap and check the speeds of 80 cars.
- a) Using the 68–95–99.7 Rule, draw and label the distribution of the proportion of these cars the police will observe speeding.
- b) Do you think the appropriate conditions necessary for your analysis are met? Explain.
12. **Smoking.** Public health statistics indicate that 26.4% of American adults smoke cigarettes. Using the 68–95–99.7 Rule, describe the sampling distribution model for the proportion of smokers among a randomly selected group of 50 adults. Be sure to discuss your assumptions and conditions.
13. **Vision.** It is generally believed that nearsightedness affects about 12% of all children. A school district has registered 170 incoming kindergarten children.
- a) Can you apply the Central Limit Theorem to describe the sampling distribution model for the sample proportion of children who are nearsighted? Check the conditions and discuss any assumptions you need to make.
- b) Sketch and clearly label the sampling model, based on the 68–95–99.7 Rule.
- c) How many of the incoming students might the school expect to be nearsighted? Explain.
14. **Mortgages.** In early 2007 the Mortgage Lenders Association reported that homeowners, hit hard by rising interest rates on adjustable-rate mortgages, were defaulting in record numbers. The foreclosure rate of 1.6% meant that millions of families were losing their homes. Suppose a large bank holds 1731 adjustable-rate mortgages.
- a) Can you apply the Central Limit Theorem to describe the sampling distribution model for the sample proportion of foreclosures? Check the conditions and discuss any assumptions you need to make.
- b) Sketch and clearly label the sampling model, based on the 68–95–99.7 Rule.
- c) How many of these homeowners might the bank expect will default on their mortgages? Explain.

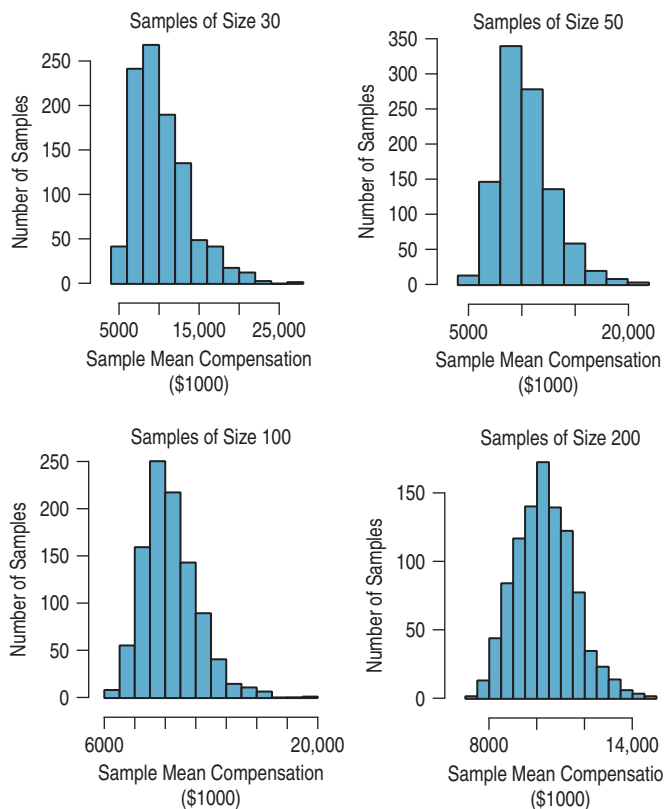
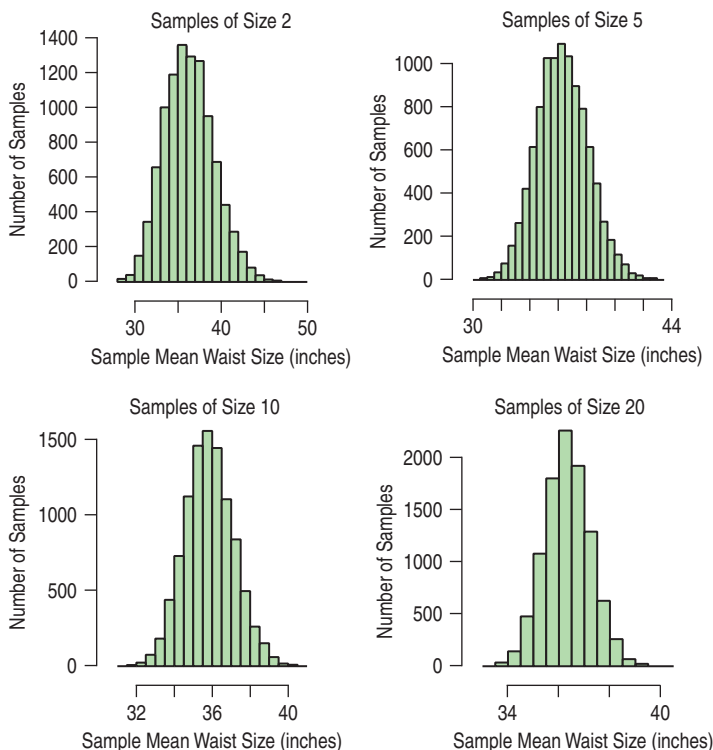


15. **Loans.** Based on past experience, a bank believes that 7% of the people who receive loans will not make payments on time. The bank has recently approved 200 loans.
- What are the mean and standard deviation of the proportion of clients in this group who may not make timely payments?
  - What assumptions underlie your model? Are the conditions met? Explain.
  - What's the probability that over 10% of these clients will not make timely payments?
16. **Contacts.** Assume that 30% of students at a university wear contact lenses.
- We randomly pick 100 students. Let  $\hat{p}$  represent the proportion of students in this sample who wear contacts. What's the appropriate model for the distribution of  $\hat{p}$ ? Specify the name of the distribution, the mean, and the standard deviation. Be sure to verify that the conditions are met.
  - What's the approximate probability that more than one third of this sample wear contacts?
17. **Back to school?** Best known for its testing program, ACT, Inc., also compiles data on a variety of issues in education. In 2004 the company reported that the national college freshman-to-sophomore retention rate held steady at 74% over the previous four years. Consider random samples of 400 freshmen who took the ACT. Use the 68–95–99.7 Rule to describe the sampling distribution model for the percentage of those students we expect to return to that school for their sophomore years. Do you think the appropriate conditions are met?
18. **Binge drinking.** As we learned in Chapter 15, a national study found that 44% of college students engage in binge drinking (5 drinks at a sitting for men, 4 for women). Use the 68–95–99.7 Rule to describe the sampling distribution model for the proportion of students in a randomly selected group of 200 college students who engage in binge drinking. Do you think the appropriate conditions are met?
19. **Back to school, again.** Based on the 74% national retention rate described in Exercise 17, does a college where 522 of the 603 freshman returned the next year as sophomores have a right to brag that it has an unusually high retention rate? Explain.
20. **Binge sample.** After hearing of the national result that 44% of students engage in binge drinking (5 drinks at a sitting for men, 4 for women), a professor surveyed a random sample of 244 students at his college and found that 96 of them admitted to binge drinking in the past week. Should he be surprised at this result? Explain.
21. **Polling.** Just before a referendum on a school budget, a local newspaper polls 400 voters in an attempt to predict whether the budget will pass. Suppose that the budget actually has the support of 52% of the voters. What's the probability the newspaper's sample will lead them to predict defeat? Be sure to verify that the assumptions and conditions necessary for your analysis are met.
22. **Seeds.** Information on a packet of seeds claims that the germination rate is 92%. What's the probability that more than 95% of the 160 seeds in the packet will germinate? Be sure to discuss your assumptions and check the conditions that support your model.
23. **Apples.** When a truckload of apples arrives at a packing plant, a random sample of 150 is selected and examined for bruises, discoloration, and other defects. The whole truckload will be rejected if more than 5% of the sample is unsatisfactory. Suppose that in fact 8% of the apples on the truck do not meet the desired standard. What's the probability that the shipment will be accepted anyway?
24. **Genetic defect.** It's believed that 4% of children have a gene that may be linked to juvenile diabetes. Researchers hoping to track 20 of these children for several years test 732 newborns for the presence of this gene. What's the probability that they find enough subjects for their study?
25. **Nonsmokers.** While some nonsmokers do not mind being seated in a smoking section of a restaurant, about 60% of the customers demand a smoke-free area. A new restaurant with 120 seats is being planned. How many seats should be in the nonsmoking area in order to be very sure of having enough seating there? Comment on the assumptions and conditions that support your model, and explain what "very sure" means to you.
26. **Meals.** A restaurateur anticipates serving about 180 people on a Friday evening, and believes that about 20% of the patrons will order the chef's steak special. How many of those meals should he plan on serving in order to be pretty sure of having enough steaks on hand to meet customer demand? Justify your answer, including an explanation of what "pretty sure" means to you.
27. **Sampling.** A sample is chosen randomly from a population that can be described by a Normal model.
- What's the sampling distribution model for the sample mean? Describe shape, center, and spread.
  - If we choose a larger sample, what's the effect on this sampling distribution model?
28. **Sampling, part II.** A sample is chosen randomly from a population that was strongly skewed to the left.
- Describe the sampling distribution model for the sample mean if the sample size is small.
  - If we make the sample larger, what happens to the sampling distribution model's shape, center, and spread?
  - As we make the sample larger, what happens to the expected distribution of the data in the sample?
29. **Waist size.** A study measured the *Waist Size* of 250 men, finding a mean of 36.33 inches and a standard deviation of 4.02 inches. Here is a histogram of these measurements

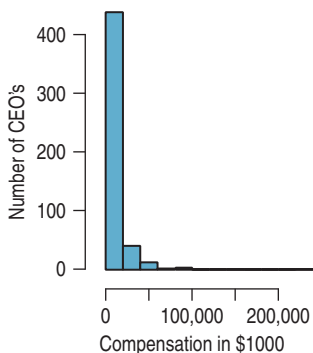


- Describe the histogram of *Waist Size*.

b) To explore how the mean might vary from sample to sample, they simulated by drawing many samples of size 2, 5, 10, and 20, with replacement, from the 250 measurements. Here are histograms of the sample means for each simulation. Explain how these histograms demonstrate what the Central Limit Theorem says about the sampling distribution model for sample means.



30. **CEO compensation.** In Chapter 5 we saw the distribution of the total compensation of the chief executive officers (CEOs) of the 800 largest U.S. companies (the Fortune 800). The average compensation (in thousands of dollars) is 10,307.31 and the standard deviation is 17,964.62. Here is a histogram of their annual compensations (in \$1000):



a) Describe the histogram of *Total Compensation*. A research organization simulated sample means by drawing samples of 30, 50, 100, and 200, with replacement, from the 800 CEOs. The histograms show the distributions of means for many samples of each size.

b) Explain how these histograms demonstrate what the Central Limit Theorem says about the sampling distribution model for sample means. Be sure to talk about shape, center, and spread.  
 c) Comment on the “rule of thumb” that “With a sample size of at least 30, the sampling distribution of the mean is Normal”?

31. **Waist size revisited.** Researchers measured the *Waist Sizes* of 250 men in a study on body fat. The true mean and standard deviation of the *Waist Sizes* for the 250 men are 36.33 in and 4.019 inches, respectively. In Exercise 29 you looked at the histograms of simulations that drew samples of sizes 2, 5, 10, and 20 (with replacement). The summary statistics for these simulations were as follows:

<i>n</i>	mean	st. dev.
2	36.314	2.855
5	36.314	1.805
10	36.341	1.276
20	36.339	0.895

a) According to the Central Limit Theorem, what should the theoretical mean and standard deviation be for each of these sample sizes?  
 b) How close are the theoretical values to what was observed in the simulation?  
 c) Looking at the histograms in Exercise 29, at what sample size would you be comfortable using the Normal model as an approximation for the sampling distribution?  
 d) What about the shape of the distribution of *Waist Size* explains your choice of sample size in part c)?

32. **CEOs revisited.** In Exercise 30 you looked at the annual compensation for 800 CEOs, for which the true mean and standard deviation were (in thousands of dollars) 10,307.31 and 17,964.62, respectively. A simulation drew samples of sizes 30, 50, 100, and 200 (with replacement) from the total annual compensations of the Fortune 800 CEOs. The summary statistics for these simulations were as follows:

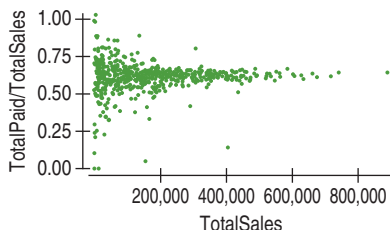
$n$	mean	st. dev.
30	10,251.73	3359.64
50	10,343.93	2483.84
100	10,329.94	1779.18
200	10,340.37	1230.79

- a) According to the Central Limit Theorem, what should the theoretical mean and standard deviation be for each of these sample sizes?
- b) How close are the theoretical values to what was observed from the simulation?
- c) Looking at the histograms in Exercise 30, at what sample size would you be comfortable using the Normal model as an approximation for the sampling distribution?
- d) What about the shape of the distribution of *Total Compensation* explains your answer in part c?

33. **GPA's.** A college's data about the incoming freshmen indicates that the mean of their high school GPAs was 3.4, with a standard deviation of 0.35; the distribution was roughly mound-shaped and only slightly skewed. The students are randomly assigned to freshman writing seminars in groups of 25. What might the mean GPA of one of these seminar groups be? Describe the appropriate sampling distribution model—shape, center, and spread—with attention to assumptions and conditions. Make a sketch using the 68–95–99.7 Rule.

34. **Home values.** Assessment records indicate that the value of homes in a small city is skewed right, with a mean of \$140,000 and standard deviation of \$60,000. To check the accuracy of the assessment data, officials plan to conduct a detailed appraisal of 100 homes selected at random. Using the 68–95–99.7 Rule, draw and label an appropriate sampling model for the mean value of the homes selected.

**T** 35. **Lucky Spot?** A reporter working on a story about the New York lottery contacted one of the authors of this book, wanting help analyzing data to see if some ticket sales outlets were more likely to produce winners. His data for each of the 966 New York lottery outlets are graphed below; the scatterplot shows the ratio *TotalPaid/TotalSales* vs. *TotalSales* for the state's "instant winner" games for all of 2007.



The reporter thinks that by identifying the outlets with the highest fraction of bets paid out, players might be able to increase their chances of winning. (Typically—but not always—instant winners are paid immediately (instantly) at the store at which they are purchased. However, the fact that tickets may be scratched off and then cashed in at any outlet may account for some outlets paying out more than they take in. The few with very low payouts may be on interstate highways where players may purchase cards but then leave.)

- a) Explain why the plot has this funnel shape.
  - b) Explain why the reporter's idea wouldn't have worked anyway.
36. **Safe cities.** Allstate Insurance Company identified the 10 safest and 10 least-safe U.S. cities from among the 200 largest cities in the United States, based on the mean number of years drivers went between automobile accidents. The cities on both lists were all smaller than the 10 largest cities. Using facts about the sampling distribution model of the mean, explain why this is not surprising.
37. **Pregnancy.** Assume that the duration of human pregnancies can be described by a Normal model with mean 266 days and standard deviation 16 days.
- a) What percentage of pregnancies should last between 270 and 280 days?
  - b) At least how many days should the longest 25% of all pregnancies last?
  - c) Suppose a certain obstetrician is currently providing prenatal care to 60 pregnant women. Let  $\bar{y}$  represent the mean length of their pregnancies. According to the Central Limit Theorem, what's the distribution of this sample mean,  $\bar{y}$ ? Specify the model, mean, and standard deviation.
  - d) What's the probability that the mean duration of these patients' pregnancies will be less than 260 days?
38. **Rainfall.** Statistics from Cornell's Northeast Regional Climate Center indicate that Ithaca, NY, gets an average of 35.4" of rain each year, with a standard deviation of 4.2". Assume that a Normal model applies.
- a) During what percentage of years does Ithaca get more than 40" of rain?
  - b) Less than how much rain falls in the driest 20% of all years?
  - c) A Cornell University student is in Ithaca for 4 years. Let  $\bar{y}$  represent the mean amount of rain for those 4 years. Describe the sampling distribution model of this sample mean,  $\bar{y}$ .
  - d) What's the probability that those 4 years average less than 30" of rain?
39. **Pregnant again.** The duration of human pregnancies may not actually follow the Normal model described in Exercise 37.
- a) Explain why it may be somewhat skewed to the left.
  - b) If the correct model is in fact skewed, does that change your answers to parts a, b, and c of Exercise 37? Explain why or why not for each.
40. **At work.** Some business analysts estimate that the length of time people work at a job has a mean of 6.2 years and a standard deviation of 4.5 years.

- a) Explain why you suspect this distribution may be skewed to the right.
- b) Explain why you could estimate the probability that 100 people selected at random had worked for their employers an average of 10 years or more, but you could not estimate the probability that an individual had done so.
41. **Dice and dollars.** You roll a die, winning nothing if the number of spots is odd, \$1 for a 2 or a 4, and \$10 for a 6.
- Find the expected value and standard deviation of your prospective winnings.
  - You play twice. Find the mean and standard deviation of your total winnings.
  - You play 40 times. What's the probability that you win at least \$100?
42. **New game.** You pay \$10 and roll a die. If you get a 6, you win \$50. If not, you get to roll again. If you get a 6 this time, you get your \$10 back.
- Create a probability model for this game.
  - Find the expected value and standard deviation of your prospective winnings.
  - You play this game five times. Find the expected value and standard deviation of your average winnings.
  - 100 people play this game. What's the probability the person running the game makes a profit?
43. **AP Stats 2006.** The College Board reported the score distribution shown in the table for all students who took the 2006 AP Statistics exam.

Score	Percent of Students
5	12.6
4	22.2
3	25.3
2	18.3
1	21.6

- Find the mean and standard deviation of the scores.
  - If we select a random sample of 40 AP Statistics students, would you expect their scores to follow a Normal model? Explain.
  - Consider the mean scores of random samples of 40 AP Statistics students. Describe the sampling model for these means (shape, center, and spread).
44. **Museum membership.** A museum offers several levels of membership, as shown in the table.

Member Category	Amount of Donation (\$)	Percent of Members
Individual	50	41
Family	100	37
Sponsor	250	14
Patron	500	7
Benefactor	1000	1

- Find the mean and standard deviation of the donations.
- During their annual membership drive, they hope to sign up 50 new members each day. Would you expect

- the distribution of the donations for a day to follow a Normal model? Explain.
- c) Consider the mean donation of the 50 new members each day. Describe the sampling model for these means (shape, center, and spread).
45. **AP Stats 2006, again.** An AP Statistics teacher had 63 students preparing to take the AP exam discussed in Exercise 43. Though they were obviously not a random sample, he considered his students to be "typical" of all the national students. What's the probability that his students will achieve an average score of at least 3?
46. **Joining the museum.** One of the museum's phone volunteers sets a personal goal of getting an average donation of at least \$100 from the new members she enrolls during the membership drive. If she gets 80 new members and they can be considered a random sample of all the museum's members, what is the probability that she can achieve her goal?
47. **Pollution.** Carbon monoxide (CO) emissions for a certain kind of car vary with mean 2.9 g/mi and standard deviation 0.4 g/mi. A company has 80 of these cars in its fleet. Let  $\bar{y}$  represent the mean CO level for the company's fleet.
- What's the approximate model for the distribution of  $\bar{y}$ ? Explain.
  - Estimate the probability that  $\bar{y}$  is between 3.0 and 3.1 g/mi.
  - There is only a 5% chance that the fleet's mean CO level is greater than what value?
48. **Potato chips.** The weight of potato chips in a medium-size bag is stated to be 10 ounces. The amount that the packaging machine puts in these bags is believed to have a Normal model with mean 10.2 ounces and standard deviation 0.12 ounces.
- What fraction of all bags sold are underweight?
  - Some of the chips are sold in "bargain packs" of 3 bags. What's the probability that none of the 3 is underweight?
  - What's the probability that the mean weight of the 3 bags is below the stated amount?
  - What's the probability that the mean weight of a 24-bag case of potato chips is below 10 ounces?
49. **Tips.** A waiter believes the distribution of his tips has a model that is slightly skewed to the right, with a mean of \$9.60 and a standard deviation of \$5.40.
- Explain why you cannot determine the probability that a given party will tip him at least \$20.
  - Can you estimate the probability that the next 4 parties will tip an average of at least \$15? Explain.
  - Is it likely that his 10 parties today will tip an average of at least \$15? Explain.
50. **Groceries.** A grocery store's receipts show that Sunday customer purchases have a skewed distribution with a mean of \$32 and a standard deviation of \$20.
- Explain why you cannot determine the probability that the next Sunday customer will spend at least \$40.
  - Can you estimate the probability that the next 10 Sunday customers will spend an average of at least \$40? Explain.
  - Is it likely that the next 50 Sunday customers will spend an average of at least \$40? Explain.

51. **More tips.** The waiter in Exercise 49 usually waits on about 40 parties over a weekend of work.
- Estimate the probability that he will earn at least \$500 in tips.
  - How much does he earn on the best 10% of such weekends?
52. **More groceries.** Suppose the store in Exercise 50 had 312 customers this Sunday.
- Estimate the probability that the store's revenues were at least \$10,000.
  - If, on a typical Sunday, the store serves 312 customers, how much does the store take in on the worst 10% of such days?
53. **IQs.** Suppose that IQs of East State University's students can be described by a Normal model with mean 130 and standard deviation 8 points. Also suppose that IQs of students from West State University can be described by a Normal model with mean 120 and standard deviation 10.
- We select a student at random from East State. Find the probability that this student's IQ is at least 125 points.
  - We select a student at random from each school. Find the probability that the East State student's IQ is at least 5 points higher than the West State student's IQ.
  - We select 3 West State students at random. Find the probability that this group's average IQ is at least 125 points.
  - We also select 3 East State students at random. What's the probability that their average IQ is at least 5 points higher than the average for the 3 West Staters?
54. **Milk.** Although most of us buy milk by the quart or gallon, farmers measure daily production in pounds. Ayrshire cows average 47 pounds of milk a day, with a standard deviation of 6 pounds. For Jersey cows, the mean daily production is 43 pounds, with a standard

deviation of 5 pounds. Assume that Normal models describe milk production for these breeds.

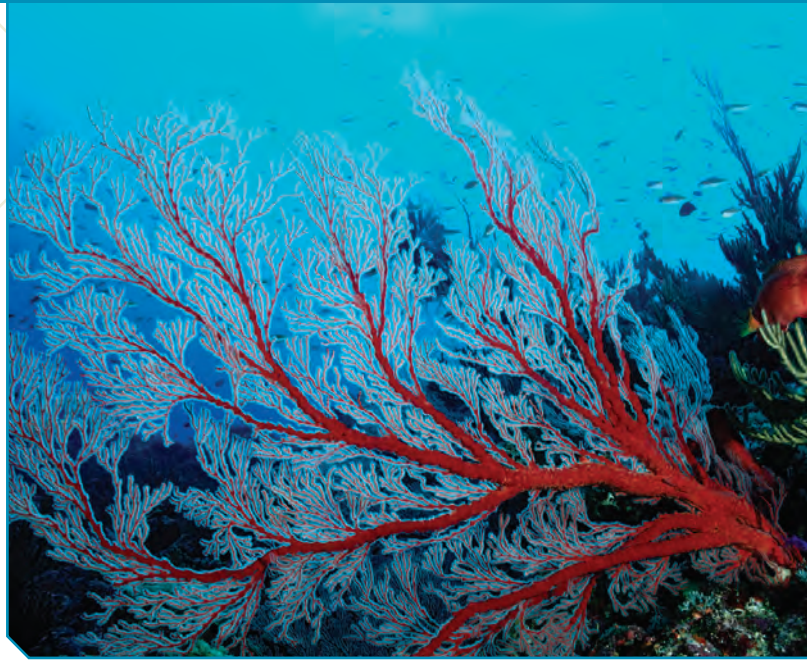
- We select an Ayrshire at random. What's the probability that she averages more than 50 pounds of milk a day?
- What's the probability that a randomly selected Ayrshire gives more milk than a randomly selected Jersey?
- A farmer has 20 Jerseys. What's the probability that the average production for this small herd exceeds 45 pounds of milk a day?
- A neighboring farmer has 10 Ayrshires. What's the probability that his herd average is at least 5 pounds higher than the average for part c's Jersey herd?



### JUST CHECKING Answers

- A Normal model (approximately).
- At the actual proportion of all students who are in favor.
- $SD(\hat{p}) = \sqrt{\frac{(0.5)(0.5)}{100}} = 0.05$
- No, this is a histogram of individuals. It may or may not be Normal, but we can't tell from the information provided.
- A Normal model (approximately).
- 266 days
- $\frac{16}{\sqrt{100}} = 1.6$  days

# Confidence Intervals for Proportions



<b>WHO</b>	Sea fans
<b>WHAT</b>	Percent infected
<b>WHEN</b>	June 2000
<b>WHERE</b>	Las Redes Reef, Akumal, Mexico, 40 feet deep
<b>WHY</b>	Research

Coral reef communities are home to one quarter of all marine plants and animals worldwide. These reefs support large fisheries by providing breeding grounds and safe havens for young fish of many species. Coral reefs are seawalls that protect shorelines against tides, storm surges, and hurricanes, and are sand “factories” that produce the limestone and sand of which beaches are made. Beyond the beach, these reefs are major tourist attractions for snorkelers and divers, driving a tourist industry worth tens of billions of dollars.

But marine scientists say that 10% of the world’s reef systems have been destroyed in recent times. At current rates of loss, 70% of the reefs could be gone in 40 years. Pollution, global warming, outright destruction of reefs, and increasing acidification of the oceans are all likely factors in this loss.

Dr. Drew Harvell’s lab studies corals and the diseases that affect them. They sampled sea fans<sup>1</sup> at 19 randomly selected reefs along the Yucatan peninsula and diagnosed whether the animals were affected by the disease *aspergillosis*.<sup>2</sup> In specimens collected at a depth of 40 feet at the Las Redes Reef in Akumal, Mexico, these scientists found that 54 of 104 sea fans sampled were infected with that disease.

Of course, we care about much more than these particular 104 sea fans. We care about the health of coral reef communities throughout the Caribbean. What can this study tell us about the prevalence of the disease among sea fans?

We have a sample proportion, which we write as  $\hat{p}$ , of 54/104, or 51.9%. Our first guess might be that this observed proportion is close to the population proportion,  $p$ . But we also know that because of natural sampling variability, if the researchers had drawn a second sample of 104 sea fans at roughly the same time, the proportion infected from that sample probably wouldn’t have been exactly 51.9%.

<sup>1</sup> That’s a sea fan in the picture. Although they look like trees, they are actually colonies of genetically identical animals.

<sup>2</sup> K. M. Mullen, C. D. Harvell, A. P. Alker, D. Dube, E. Jordán-Dahlgren, J. R. Ward, and L. E. Petes, “Host range and resistance to aspergillosis in three sea fan species from the Yucatan,” *Marine Biology* (2006), Springer-Verlag.

What *can* we say about the population proportion,  $p$ ? To start to answer this question, think about how different the sample proportion might have been if we'd taken another random sample from the same population. But wait. Remember—we aren't actually going to take more samples. We just want to *imagine* how the sample proportions might vary from sample to sample. In other words, we want to know about the *sampling distribution* of the sample proportion of infected sea fans.

## A Confidence Interval

**AS** **Activity: Confidence Intervals and Sampling Distributions.** Simulate the sampling distribution, and see how it gives a confidence interval.

**NOTATION ALERT:**

Remember that  $\hat{p}$  is our sample-based estimate of the true proportion  $p$ . Recall also that  $q$  is just shorthand for  $1 - p$ , and  $\hat{q} = 1 - \hat{p}$ .

When we use  $\hat{p}$  to estimate the standard deviation of the sampling distribution model, we call that the **standard error** and write  $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$

Let's look at our model for the sampling distribution. What do we know about it? We know it's approximately Normal (under certain assumptions, which we should be careful to check) and that its mean is the proportion of all infected sea fans on the Las Redes Reef. Is the infected proportion of *all* sea fans 51.9%? No, that's just  $\hat{p}$ , our estimate. We don't know the proportion,  $p$ , of all the infected sea fans; that's what we're trying to find out. We do know, though, that the sampling distribution model of  $\hat{p}$  is centered at  $p$ , and we know that the standard deviation of the sampling distribution is  $\sqrt{\frac{pq}{n}}$ .

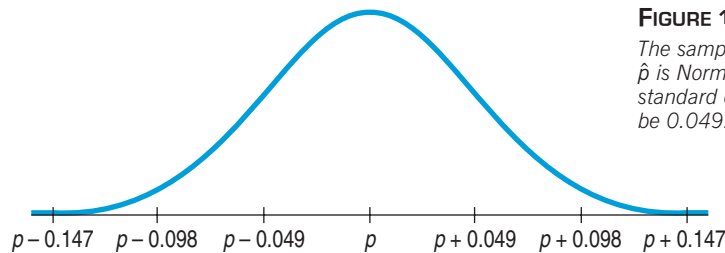
Now we have a problem: Since we don't know  $p$ , we can't find the true standard deviation of the sampling distribution model. We do know the observed proportion,  $\hat{p}$ , so, of course we just use what we know, and we estimate. That may not seem like a big deal, but it gets a special name. **Whenever we estimate the standard deviation of a sampling distribution, we call it a **standard error**.**<sup>3</sup> For a sample proportion,  $\hat{p}$ , the standard error is

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

For the sea fans, then:

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.519)(0.481)}{104}} = 0.049 = 4.9\%$$

Now we know that the sampling model for  $\hat{p}$  should look like this:



**FIGURE 19.1**  
The sampling distribution model for  $\hat{p}$  is Normal with a mean of  $p$  and a standard deviation we estimate to be 0.049.

Great. What does that tell us? Well, because it's Normal, it says that about 68% of all samples of 104 sea fans will have  $\hat{p}$ 's within 1 SE, 0.049, of  $p$ . And about 95% of all these samples will be within  $p \pm 2$  SEs. But where is *our* sample proportion in this picture? And what value does  $p$  have? We still don't know!

We do know that for 95% of random samples,  $\hat{p}$  will be no more than 2 SEs away from  $p$ . So let's look at this from  $\hat{p}$ 's point of view. If I'm  $\hat{p}$ , there's a 95%

<sup>3</sup>This isn't such a great name because it isn't standard and nobody made an error. But it's much shorter and more convenient than saying, "the estimated standard deviation of the sampling distribution of the sample statistic."

chance that  $p$  is no more than 2 SEs away from me. If I reach out 2 SEs, or  $2 \times 0.049$ , away from me on both sides, I'm 95% sure that  $p$  will be within my grasp. Now I've got him! Probably. Of course, even if my interval does catch  $p$ , I still don't know its true value. The best I can do is an interval, and even then I can't be positive it contains  $p$ .

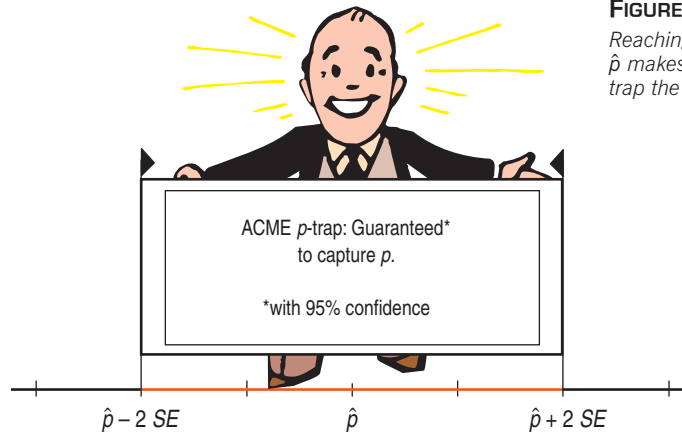


FIGURE 19.2

Reaching out 2 SEs on either side of  $\hat{p}$  makes us 95% confident that we'll trap the true proportion,  $p$ .

So what can we really say about  $p$ ? Here's a list of things we'd like to be able to say, in order of strongest to weakest and the reasons we can't say most of them:

**A S** **Activity: Can We Estimate a Parameter?** Consider these four interpretations of a confidence interval by simulating to see whether they could be right.

*"Far better an approximate answer to the right question, . . . than an exact answer to the wrong question."*

—John W. Tukey

1. **"51.9% of all sea fans on the Las Redes Reef are infected."** It would be nice to be able to make absolute statements about population values with certainty, but we just don't have enough information to do that. There's no way to be sure that the population proportion is the same as the sample proportion; in fact, it almost certainly isn't. Observations vary. Another sample would yield a different sample proportion.
2. **"It is probably true that 51.9% of all sea fans on the Las Redes Reef are infected."** No. In fact, we can be pretty sure that whatever the true proportion is, it's not exactly 51.900%. So the statement is not true.
3. **"We don't know exactly what proportion of sea fans on the Las Redes Reef is infected, but we know that it's within the interval 51.9%  $\pm$  2  $\times$  4.9%. That is, it's between 42.1% and 61.7%."** This is getting closer, but we still can't be certain. We can't know *for sure* that the true proportion is in this interval—or in any particular interval.
4. **"We don't know exactly what proportion of sea fans on the Las Redes Reef is infected, but the interval from 42.1% to 61.7% probably contains the true proportion."** We've now fudged twice—first by giving an interval and second by admitting that we only think the interval "probably" contains the true value. And this statement is true.

That last statement may be true, but it's a bit wishy-washy. We can tighten it up a bit by quantifying what we mean by "probably." We saw that 95% of the time when we reach out 2 SEs from  $\hat{p}$  we capture  $p$ , so we can be 95% confident that this is one of those times. After putting a number on the probability that this interval covers the true proportion, we've given our best guess of where the parameter is and how certain we are that it's within some range.

5. **"We are 95% confident that between 42.1% and 61.7% of Las Redes sea fans are infected."** Statements like these are called **confidence intervals**. They're the best we can do.

Each confidence interval discussed in the book has a name. You'll see many different kinds of confidence intervals in the following chapters. Some will be



about more than *one* sample, some will be about statistics other than *proportions*, and some will use models other than the Normal. The interval calculated and interpreted here is sometimes called a **one-proportion z-interval**.<sup>4</sup>



### JUST CHECKING

A Pew Research study regarding cell phones asked questions about cell phone experience. One growing concern is unsolicited advertising in the form of text messages. Pew asked cell phone owners, “Have you ever received unsolicited text messages on your cell phone from advertisers?” and 17% reported that they had. Pew estimates a 95% confidence interval to be  $0.17 \pm 0.04$ , or between 13% and 21%.

Are the following statements about people who have cell phones correct? Explain.

1. In Pew’s sample, somewhere between 13% and 21% of respondents reported that they had received unsolicited advertising text messages.
2. We can be 95% confident that 17% of U.S. cell phone owners have received unsolicited advertising text messages.
3. We are 95% confident that between 13% and 21% of all U.S. cell phone owners have received unsolicited advertising text messages.
4. We know that between 13% and 21% of all U.S. cell phone owners have received unsolicited advertising text messages.
5. 95% of all U.S. cell phone owners have received unsolicited advertising text messages.

## What Does “95% Confidence” Really Mean?

What do we mean when we say we have 95% confidence that our interval contains the true proportion? Formally, what we mean is that “95% of samples of this size will produce confidence intervals that capture the true proportion.” This is correct, but a little long winded, so we sometimes say, “we are 95% confident that the true proportion lies in our interval.” Our uncertainty is about whether the particular sample we have at hand is one of the successful ones or one of the 5% that fail to produce an interval that captures the true value.

Back in Chapter 18 we saw that proportions vary from sample to sample. If other researchers select their own samples of sea fans, they’ll also find some infected by the disease, but each person’s sample proportion will almost certainly differ from ours. When they each try to estimate the true rate of infection in the entire population, they’ll center *their* confidence intervals at the proportions they observed in their own samples. Each of us will end up with a different interval.

Our interval guessed the true proportion of infected sea fans to be between about 42% and 62%. Another researcher whose sample contained more infected fans than ours did might guess between 46% and 66%. Still another who happened to collect fewer infected fans might estimate the true proportion to be between 23% and 43%. And so on. Every possible sample would produce yet another confidence interval. Although wide intervals like these can’t pin down the actual rate of infection very precisely, we expect that most of them should be winners, capturing the true value. Nonetheless, some will be duds, missing the population proportion entirely.

On the next page you’ll see confidence intervals produced by simulating 20 different random samples. The red dots are the proportions of infected fans in

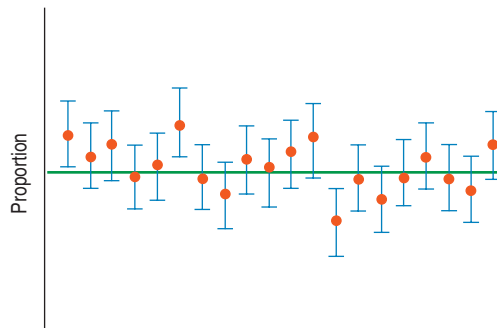
**A S** **Activity: Confidence Intervals for Proportions.** This new interactive tool makes it easy to construct and experiment with confidence intervals. We’ll use this tool for the rest of the course—sure beats calculating by hand!

<sup>4</sup> In fact, this confidence interval is so standard for a single proportion that you may see it simply called a “confidence interval for the proportion.”

**TI-*nspire***

**Confidence intervals.** Generate confidence intervals from many samples to see how often they capture the true proportion.

each sample, and the blue segments show the confidence intervals found for each. The green line represents the true rate of infection in the population, so you can see that most of the intervals caught it—but a few missed. (And notice again that it is the *intervals* that vary from sample to sample; the green line doesn't move.)



The horizontal green line shows the true percentage of all sea fans that are infected. Most of the 20 simulated samples produced confidence intervals that captured the true value, but a few missed.

Of course, there's a huge number of possible samples that *could* be drawn, each with its own sample proportion. These are just some of them. Each sample proportion can be used to make a confidence interval. That's a large pile of possible confidence intervals, and ours is just one of those in the pile. Did *our* confidence interval "work"? We can never be sure, because we'll never know the true proportion of all the sea fans that are infected. However, the Central Limit Theorem assures us that 95% of the intervals in the pile are winners, covering the true value, and only 5% are duds. *That's* why we're 95% confident that our interval is a winner!

**FOR EXAMPLE****Polls and margin of error**

On January 30–31, 2007, Fox News/Opinion Dynamics polled 900 registered voters nationwide.<sup>5</sup> When asked, "Do you believe global warming exists?" 82% said "Yes". Fox reported their margin of error to be  $\pm 3\%$ .

**Question:** It is standard among pollsters to use a 95% confidence level unless otherwise stated. Given that, what does Fox News mean by claiming a margin of error of  $\pm 3\%$  in this context?

If this polling were done repeatedly, 95% of all random samples would yield estimates that come within  $\pm 3\%$  of the true proportion of all registered voters who believe that global warming exists.

## Margin of Error: Certainty vs. Precision

We've just claimed that with a certain confidence we've captured the true proportion of all infected sea fans. Our confidence interval had the form

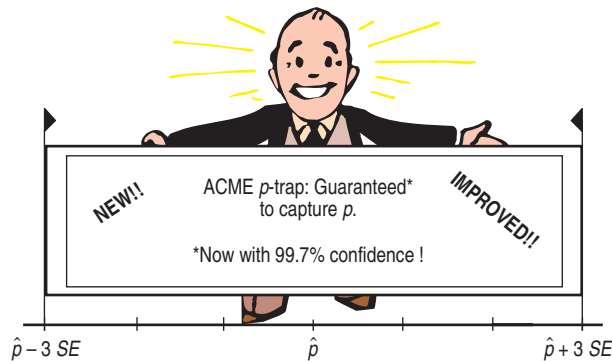
$$\hat{p} \pm 2 SE(\hat{p}).$$

The extent of the interval on either side of  $\hat{p}$  is called the **margin of error (ME)**. We'll want to use the same approach for many other situations besides estimating proportions. In general, confidence intervals look like this:

$$\text{Estimate} \pm ME.$$

<sup>5</sup> www.foxnews.com, "Fox News Poll: Most Americans Believe in Global Warming," Feb 7, 2007.

The margin of error for our 95% confidence interval was 2 SE. What if we wanted to be more confident? To be more confident, we'll need to capture  $p$  more often, and to do that we'll need to make the interval wider. For example, if we want to be 99.7% confident, the margin of error will have to be 3 SE.



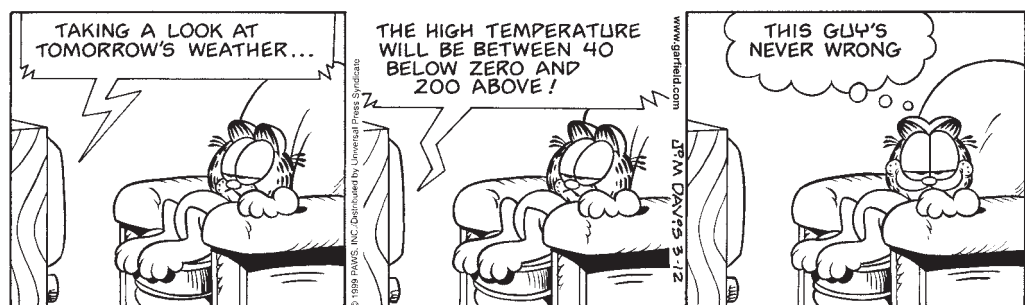
**FIGURE 19.3**

Reaching out 3 SEs on either side of  $\hat{p}$  makes us 99.7% confident we'll trap the true proportion  $p$ . Compare with Figure 19.2.

**A S** **Activity: Balancing Precision and Certainty.** What percent of parents expect their kids to pay for college with a student loan? Investigate the balance between the precision and the certainty of a confidence interval.

The more confident we want to be, the larger the margin of error must be. We can be 100% confident that the proportion of infected sea fans is between 0% and 100%, but this isn't likely to be very useful. On the other hand, we could give a confidence interval from 51.8% to 52.0%, but we can't be very confident about a precise statement like this. Every confidence interval is a balance between certainty and precision.

The tension between certainty and precision is always there. Fortunately, in most cases we can be both sufficiently certain and sufficiently precise to make useful statements. There is no simple answer to the conflict. You must choose a confidence level yourself. The data can't do it for you. The choice of confidence level is somewhat arbitrary. The most commonly chosen confidence levels are 90%, 95%, and 99%, but any percentage can be used. (In practice, though, using something like 92.9% or 97.2% is likely to make people think you're up to something.)



Garfield © 1999 Paws, Inc. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

## FOR EXAMPLE

### Finding the margin of error (Take 1)

**Recap:** A January 2007 Fox poll of 900 registered voters reported a margin of error of  $\pm 3\%$ . It is a convention among pollsters to use a 95% confidence level and to report the “worst case” margin of error, based on  $p = 0.5$ .

**Question:** How did Fox calculate their margin of error?

$$\text{Assuming } p = 0.5, \text{ for random samples of } n = 900, SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.5)(0.5)}{900}} = 0.0167$$

For a 95% confidence level,  $ME = 2(0.0167) = 0.0333$ , so Fox's margin of error is just a bit over  $\pm 3\%$ .

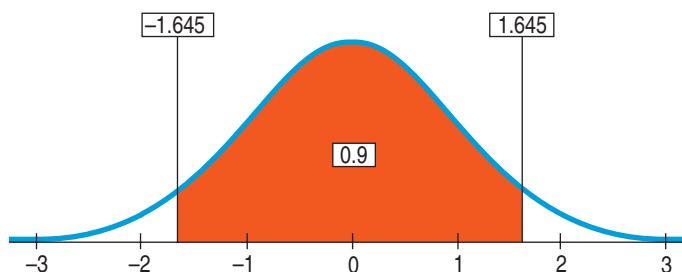
## Critical Values

### NOTATION ALERT:

We'll put an asterisk on a letter to indicate a critical value, so  $z^*$  is always a critical value from a Normal model.

In our sea fans example we used  $2SE$  to give us a 95% confidence interval. To change the confidence level, we'd need to change the *number* of SEs so that the size of the margin of error corresponds to the new level. This number of SEs is called the **critical value**. Here it's based on the Normal model, so we denote it  $z^*$ . For any confidence level, we can find the corresponding critical value from a computer, a calculator, or a Normal probability table, such as Table Z.

For a 95% confidence interval, you'll find the precise critical value is  $z^* = 1.96$ . That is, 95% of a Normal model is found within  $\pm 1.96$  standard deviations of the mean. We've been using  $z^* = 2$  from the 68–95–99.7 Rule because it's easy to remember.



**FIGURE 19.4**

For a 90% confidence interval, the critical value is 1.645, because, for a Normal model, 90% of the values are within 1.645 standard deviations from the mean.

### FOR EXAMPLE

#### Finding the margin of error (Take 2)

**Recap:** In January 2007 a Fox News poll of 900 registered voters found that 82% of the respondents believed that global warming exists. Fox reported a 95% confidence interval with a margin of error of  $\pm 3\%$ .

**Questions:** Using the critical value of  $z$  and the standard error based on the observed proportion, what would be the margin of error for a 90% confidence interval? What's good and bad about this change?

$$\text{With } n = 900 \text{ and } \hat{p} = 0.82, SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.82)(0.18)}{900}} = 0.0128$$

For a 90% confidence level,  $z^* = 1.645$ , so  $ME = 1.645(0.0128) = 0.021$

Now the margin of error is only about  $\pm 2\%$ , producing a narrower interval. That makes for a more precise estimate of voter belief, but provides less certainty that the interval actually contains the true proportion of voters believing in global warming.



### JUST CHECKING

Think some more about the 95% confidence interval Fox News created for the proportion of registered voters who believe that global warming exists.

6. If Fox wanted to be 98% confident, would their confidence interval need to be wider or narrower?
7. Fox's margin of error was about  $\pm 3\%$ . If they reduced it to  $\pm 2\%$ , would their level of confidence be higher or lower?
8. If Fox News had polled more people, would the interval's margin of error have been larger or smaller?

## Assumptions and Conditions

We've just made some pretty sweeping statements about sea fans. Those statements were possible because we used a Normal model for the sampling distribution. But is that model appropriate?

As we've seen, all statistical models make assumptions. Different models make different assumptions. If those assumptions are not true, the model might be inappropriate and our conclusions based on it may be wrong. Because the confidence interval is built on the Normal model for the sampling distribution, the assumptions and conditions are the same as those we discussed in Chapter 18. But, because they are so important, we'll go over them again.

We can never be certain that an assumption is true, but we can decide intelligently whether it is reasonable. When we have data, we can often decide whether an assumption is plausible by checking a related condition. However, we want to make a statement about the world at large, not just about the data we collected. So the assumptions we make are not just about how our data look, but about how representative they are.

**AS** **Activity: Assumptions and Conditions.** Here's an animated review of the assumptions and conditions.

### INDEPENDENCE ASSUMPTION

**Independence Assumption:** We first need to *Think* about whether the independence assumption is plausible. We often look for reasons to suspect that it fails. We wonder whether there is any reason to believe that the data values somehow affect each other. (For example, might the disease in sea fans be contagious?) Whether you decide that the **Independence Assumption** is plausible depends on your knowledge of the situation. It's not one you can check by looking at the data.

However, now that we have data, there are two conditions that we can check:

**Randomization Condition:** Were the data sampled at random or generated from a properly randomized experiment? Proper randomization can help ensure independence.

**10% Condition:** Samples are almost always drawn without replacement. Usually, of course, we'd like to have as large a sample as we can. But when the population itself is small we have another concern. When we sample from small populations, the probability of success may be different for the last few individuals we draw than it was for the first few. For example, if most of the women have already been sampled, the chance of drawing a woman from the remaining population is lower. If the sample exceeds 10% of the population, the probability of a success changes so much during the sampling that our Normal model may no longer be appropriate. But if less than 10% of the population is sampled, the effect on independence is negligible.

### SAMPLE SIZE ASSUMPTION

The model we use for inference is based on the Central Limit Theorem. The **Sample Size Assumption** addresses the question of whether the sample is large enough to make the sampling model for the sample proportions approximately Normal. It turns out that we need more data as the proportion gets closer and closer to either extreme (0 or 1). We can check this assumption with the:

**Success/Failure Condition:** We must expect at least 10 "successes" and at least 10 "failures." Recall that by tradition we arbitrarily label one alternative (usually the outcome being counted) as a "success" even if it's something bad (like a sick sea fan). The other alternative is, of course, then a "failure."

## AS

**Activity: A Confidence**

**Interval for  $p$ .** View the video story of pollution in Chesapeake Bay, and make a confidence interval for the analysis with the interactive tool.

**ONE-PROPORTION  $z$ -INTERVAL**

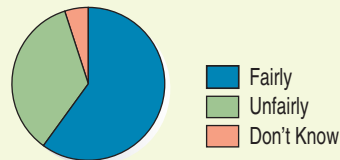
When the conditions are met, we are ready to find the confidence interval for the population proportion,  $p$ . The confidence interval is  $\hat{p} \pm z^* \times SE(\hat{p})$

where the standard deviation of the proportion is estimated by  $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ .

**STEP-BY-STEP EXAMPLE****A Confidence Interval for a Proportion**

<b>WHO</b>	Adults in the United States
<b>WHAT</b>	Response to a question about the death penalty
<b>WHEN</b>	May 2006
<b>WHERE</b>	United States
<b>HOW</b>	510 adults were randomly sampled and asked by the Gallup Poll
<b>WHY</b>	Public opinion research

In May 2006, the Gallup Poll<sup>6</sup> asked 510 randomly sampled adults the question “Generally speaking, do you believe the death penalty is applied fairly or unfairly in this country today?” Of these, 60% answered “Fairly,” 35% said “Unfairly,” and 4% said they didn’t know.



**Question:** From this survey, what can we conclude about the opinions of all adults?

To answer this question, we’ll build a confidence interval for the proportion of all U.S. adults who believe the death penalty is applied fairly. There are four steps to building a confidence interval for proportions: Plan, Model, Mechanics, and Conclusion.



**Plan** State the problem and the  $W$ 's.

Identify the *parameter* you wish to estimate.

Identify the *population* about which you wish to make statements.

Choose and state a confidence level.

**Model** Think about the assumptions and check the conditions.

I want to find an interval that is likely, with 95% confidence, to contain the true proportion,  $p$ , of U.S. adults who think the death penalty is applied fairly. I have a random sample of 510 U.S. adults.

✓ **Independence Assumption:** Gallup phoned a random sample of U.S. adults. It is very unlikely that any of their respondents influenced each other.

✓ **Randomization Condition:** Gallup drew a random sample from all U.S. adults. I don’t have details of their randomization but assume that I can trust it.

✓ **10% Condition:** Although sampling was necessarily without replacement, there are many more U.S. adults than were sampled. The sample is certainly less than 10% of the population.

<sup>6</sup> www.gallup.com

State the sampling distribution model for the statistic.

Choose your method.

✓ **Success/Failure Condition:**

$$n\hat{p} = 510(60\%) = 306 \geq 10 \text{ and}$$

$$n\hat{q} = 510(40\%) = 204 \geq 10,$$

so the sample appears to be large enough to use the Normal model.

The conditions are satisfied, so I can use a Normal model to find a **one-proportion z-interval**.



**Mechanics** Construct the confidence interval.

First find the standard error. (Remember: It's called the "standard error" because we don't know  $p$  and have to use  $\hat{p}$  instead.)

Next find the margin of error. We could informally use 2 for our critical value, but 1.96 is more accurate.

Write the confidence interval (CI).



The CI is centered at the sample proportion and about as wide as we might expect for a sample of 500.

$$n = 510, \hat{p} = 0.60, \text{ so}$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.60)(0.40)}{510}} = 0.022$$

Because the sampling model is Normal, for a 95% confidence interval, the critical value  $z^* = 1.96$ .

The margin of error is

$$ME = z^* \times SE(\hat{p}) = 1.96(0.022) = 0.043$$

So the 95% confidence interval is

$$0.60 \pm 0.043 \text{ or } (0.557, 0.643)$$



**Conclusion** Interpret the confidence interval in the proper context. We're 95% confident that our interval captured the true proportion.

I am 95% confident that between 55.7% and 64.3% of all U.S. adults think that the death penalty is applied fairly.

### TI Tips

### Finding confidence intervals

```
EDIT CALC TESTS
7:1-PropZInt...
8:TInterval...
9:2-SampZInt...
0:2-SampTInt...
1:1-PropZInt...
2:2-PropZInt...
3:4-Test...
```

```
1-PropZInt
x:54
n:104
C-Level: .95
Calculate
```

It will come as no surprise that your TI can calculate a confidence interval for a population proportion. Remember the sea fans? Of 104 sea fans, 54 were diseased. To find the resulting confidence interval, we first take a look at a whole new menu.

- Under **STAT** go to the **TESTS** menu. Quite a list! Commands are found here for the inference procedures you will learn through the coming chapters.
- We're using a Normal model to find a confidence interval for a proportion based on one sample. Scroll down the list and select **A: 1-PropZInt**.
- Enter the number of successes observed and the sample size.
- Specify a confidence level and then **Calculate**.

```
1-PropZInt
(.42321, .61525)
p̂=.5192307692
n=104
```

```
ERR:DOMAIN
Quit
```

And there it is! Note that the TI calculates the sample proportion for you, but the important result is the interval itself, 42% to 62%. The calculator did the easy part—just Show. Tell is harder. It's your job to interpret that interval correctly.

Beware: You may run into a problem. When you enter the value of  $\times$ , you need a *count*, not a percentage. Suppose the marine scientists had reported that 52% of the 104 sea fans were infected. You can enter  $\times: .52*104$ , and the calculator will evaluate that as 54.08. Wrong. Unless you fix that result, you'll get an error message. Think about it—the number of infected sea fans must have been a whole number, evidently 54. When the scientists reported the results, they rounded off the actual percentage ( $54 \div 104 = 51.923\%$ ) to 52%. Simply change the value of  $\times$  to 54 and you should be able to **Calculate** the correct interval.

## CHOOSING YOUR SAMPLE SIZE

The question of how large a sample to take is an important step in planning any study. We weren't ready to make that calculation when we first looked at study design in Chapter 12, but now we can—and we always should.

Suppose a candidate is planning a poll and wants to estimate voter support within 3% with 95% confidence. How large a sample does she need?

Let's look at the margin of error:

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.03 = 1.96 \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

We want to find  $n$ , the sample size. To find  $n$  we need a value for  $\hat{p}$ . We don't know  $\hat{p}$  because we don't have a sample yet, but we can probably guess a value. The worst case—the value that makes  $\hat{p}\hat{q}$  (and therefore  $n$ ) largest—is 0.50, so if we use that value for  $\hat{p}$ , we'll certainly be safe. Our candidate probably expects to be near 50% anyway.

Our equation, then, is

$$0.03 = 1.96 \sqrt{\frac{(0.5)(0.5)}{n}}$$

To solve for  $n$ , we first multiply both sides of the equation by  $\sqrt{n}$  and then divide by 0.03:

$$0.03\sqrt{n} = 1.96\sqrt{(0.5)(0.5)}$$

$$\sqrt{n} = \frac{1.96\sqrt{(0.5)(0.5)}}{0.03} \approx 32.67$$

Notice that evaluating this expression tells us the *square root* of the sample size. We need to square that result to find  $n$ :

$$n \approx (32.67)^2 \approx 1067.1$$

To be safe, we round up and conclude that we need at least 1068 respondents to keep the margin of error as small as 3% with a confidence level of 95%.

### What do I use instead of $\hat{p}$ ?

Often we have an estimate of the population proportion based on experience or perhaps a previous study. If so, use that value as  $\hat{p}$  in calculating what size sample you need. If not, the cautious approach is to use  $p = 0.5$  in the sample size calculation; that will determine the largest sample necessary regardless of the true proportion.



## FOR EXAMPLE

## Choosing a sample size

**Recap:** The Fox News poll which estimated that 82% of all voters believed global warming exists had a margin of error of  $\pm 3\%$ . Suppose an environmental group planning a follow-up survey of voters' opinions on global warming wants to determine a 95% confidence interval with a margin of error of no more than  $\pm 2\%$ .

**Question:** How large a sample do they need? Use the Fox News estimate as the basis for your calculation.

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$0.02 = 1.96 \sqrt{\frac{(0.82)(0.18)}{n}}$$

$$\sqrt{n} = \frac{1.96 \sqrt{(0.82)(0.18)}}{0.02} \approx 37.65$$

$$n = 37.65^2 = 1,417.55$$

The environmental group's survey will need about 1,418 respondents.

Public opinion polls often sample 1000 people, which gives an ME of 3% when  $p = 0.5$ . But businesses and nonprofit organizations typically use much larger samples to estimate the proportion who will accept a direct mail offer. Why? Because that proportion is very low—often far below 5%. An ME of 3% wouldn't be precise enough. An ME like 0.1% would be more useful, and that requires a very large sample size.

Unfortunately, bigger samples cost more money and more effort. Because the standard error declines only with the *square root* of the sample size, to cut the standard error (and thus the ME) in half, we must *quadruple* the sample size.

Generally a margin of error of 5% or less is acceptable, but different circumstances call for different standards. For a pilot study, a margin of error of 10% may be fine, so a sample of 100 will do quite well. In a close election, a polling organization might want to get the margin of error down to 2%. Drawing a large sample to get a smaller ME, however, can run into trouble. It takes time to survey 2400 people, and a survey that extends over a week or more may be trying to hit a target that moves during the time of the survey. An important event can change public opinion in the middle of the survey process.

Keep in mind that the sample size for a survey is the number of respondents, not the number of people to whom questionnaires were sent or whose phone numbers were dialed. And keep in mind that a low response rate turns any study essentially into a voluntary response study, which is of little value for inferring population values. It's almost always better to spend resources on increasing the response rate than on surveying a larger group. A full or nearly full response by a modest-size sample can yield useful results.

Surveys are not the only place where proportions pop up. Banks sample huge mailing lists to estimate what proportion of people will accept a credit card offer. Even pilot studies may mail offers to over 50,000 customers. Most don't respond; that doesn't make the sample smaller—they simply said "No thanks". Those who do respond want the card. To the bank, the response rate<sup>7</sup> is  $\hat{p}$ . With a typical success rate around 0.5%, the bank needs a very small margin of error—often as low as 0.1%—to make a sound business decision. That calls for a large sample, and the bank must take care in estimating the size needed. For our election poll calculation we used  $p = 0.5$ , both because it's safe and because we honestly believed  $p$  to be near 0.5. If the bank used 0.5, they'd get an absurd answer. Instead, they base their calculation on a proportion closer to the one they expect to find.

<sup>7</sup>In marketing studies every mailing yields a response—"yes" or "no"—and "response rate" means the proportion of customers who accept an offer. That's not the way we use the term for survey response.

## FOR EXAMPLE

## Sample size revisited

A credit card company is about to send out a mailing to test the market for a new credit card. From that sample, they want to estimate the true proportion of people who will sign up for the card nationwide. A pilot study suggests that about 0.5% of the people receiving the offer will accept it.

**Question:** To be within a tenth of a percentage point (0.001) of the true rate with 95% confidence, how big does the test mailing have to be?

$$\begin{aligned} \text{Using the estimate } \hat{p} = 0.5\%: \quad ME = 0.001 &= z^* \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \sqrt{\frac{(0.005)(0.995)}{n}} \\ (0.001)^2 &= 1.96^2 \frac{(0.005)(0.995)}{n} \Rightarrow n = \frac{1.96^2(0.005)(0.995)}{(0.001)^2} \\ &= 19,111.96 \text{ or } 19,112 \end{aligned}$$

That's a lot, but it's actually a reasonable size for a trial mailing such as this. Note, however, that if they had assumed 0.50 for the value of  $p$ , they would have found

$$\begin{aligned} ME = 0.001 &= z^* \sqrt{\frac{pq}{n}} = 1.96 \sqrt{\frac{(0.5)(0.5)}{n}} \\ (0.001)^2 &= 1.96^2 \frac{(0.5)(0.5)}{n} \Rightarrow n = \frac{1.96^2(0.5)(0.5)}{(0.001)^2} = 960,400. \end{aligned}$$

Quite a different (and unreasonable) result.

## WHAT CAN GO WRONG?

Confidence intervals are powerful tools. Not only do they tell what we know about the parameter value, but—more important—they also tell what we *don't* know. In order to use confidence intervals effectively, you must be clear about what you say about them.

### DON'T MISSTATE WHAT THE INTERVAL MEANS

- ▶ **Don't suggest that the parameter varies.** A statement like "There is a 95% chance that the true proportion is between 42.7% and 51.3%" sounds as though you think the population proportion wanders around and sometimes happens to fall between 42.7% and 51.3%. When you interpret a confidence interval, make it clear that *you* know that the population parameter is fixed and that it is the interval that varies from sample to sample.
- ▶ **Don't claim that other samples will agree with yours.** Keep in mind that the confidence interval makes a statement about the true population proportion. An interpretation such as "In 95% of samples of U.S. adults, the proportion who think marijuana should be decriminalized will be between 42.7% and 51.3%" is just wrong. The interval isn't about sample proportions but about the population proportion.
- ▶ **Don't be certain about the parameter.** Saying "Between 42.1% and 61.7% of sea fans are infected" asserts that the population proportion cannot be outside that interval. Of course, we can't be absolutely certain of that. (Just pretty sure.)
- ▶ **Don't forget: It's about the parameter.** Don't say, "I'm 95% confident that  $\hat{p}$  is between 42.1% and 61.7%." Of course you are—in fact, we calculated that  $\hat{p} = 51.9\%$  of the

(continued)

**What Can I Say?**

Confidence intervals are based on random samples, so the interval is random, too. The CLT tells us that 95% of the random samples will yield intervals that capture the true value. That's what we mean by being 95% confident.

Technically, we should say, "I am 95% confident that the interval from 42.1% to 61.7% captures the true proportion of infected sea fans." That formal phrasing emphasizes that *our confidence (and our uncertainty) is about the interval, not the true proportion*. But you may choose a more casual phrasing like "I am 95% confident that between 42.1% and 61.7% of the Las Redes fans are infected." Because you've made it clear that the uncertainty is yours and you didn't suggest that the randomness is in the true proportion, this is OK. Keep in mind that it's the interval that's random and is the focus of both our confidence and doubt.

fans in our sample were infected. So we already *know* the sample proportion. The confidence interval is about the (unknown) population parameter,  $p$ .

- ▶ **Don't claim to know too much.** Don't say, "I'm 95% confident that between 42.1% and 61.7% of all the sea fans in the world are infected." You didn't sample from all 500 species of sea fans found in coral reefs around the world. Just those of this type on the Las Redes Reef.
- ▶ **Do take responsibility.** Confidence intervals are about *uncertainty*. You are the one who is uncertain, not the parameter. You have to accept the responsibility and consequences of the fact that not all the intervals you compute will capture the true value. In fact, about 5% of the 95% confidence intervals you find will fail to capture the true value of the parameter. You *can* say, "I am 95% confident that between 42.1% and 61.7% of the sea fans on the Las Redes Reef are infected."<sup>8</sup>
- ▶ **Do treat the whole interval equally.** Although a confidence interval is a set of plausible values for the parameter, don't think that the values in the middle of a confidence interval are somehow "more plausible" than the values near the edges. Your interval provides no information about where in your current interval (if at all) the parameter value is most likely to be hiding.

**MARGIN OF ERROR TOO LARGE TO BE USEFUL**

We know we can't be exact, but how precise do we need to be? A confidence interval that says that the percentage of infected sea fans is between 10% and 90% wouldn't be of much use. Most likely, you have some sense of how large a margin of error you can tolerate. What can you do?

One way to make the margin of error smaller is to reduce your level of confidence. But that may not be a useful solution. It's a rare study that reports confidence levels lower than 80%. Levels of 95% or 99% are more common.

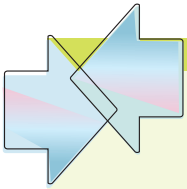
The time to think about whether your margin of error is small enough to be useful is when you design your study. Don't wait until you compute your confidence interval. To get a narrower interval without giving up confidence, you need to have less variability in your sample proportion. How can you do that? Choose a larger sample.

**VIOLATIONS OF ASSUMPTIONS**

Confidence intervals and margins of error are often reported along with poll results and other analyses. But it's easy to misuse them and wise to be aware of the ways things can go wrong.

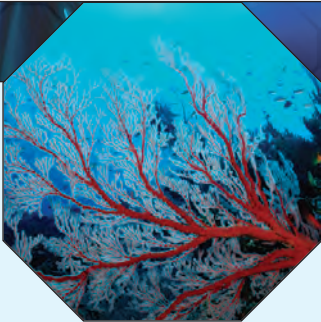
- ▶ **Watch out for biased sampling.** Don't forget about the potential sources of bias in surveys that we discussed in Chapter 12. Just because we have more statistical machinery now doesn't mean we can forget what we've already learned. A questionnaire that finds that 85% of people enjoy filling out surveys still suffers from nonresponse bias even though now we're able to put confidence intervals around this (biased) estimate.
- ▶ **Think about independence.** The assumption that the values in our sample are mutually independent is one that we usually cannot check. It always pays to think about it, though. For example, the disease affecting the sea fans might be contagious, so that fans growing near a diseased fan are more likely themselves to be diseased. Such contagion would violate the Independence Assumption and could severely affect our sample proportion. It could be that the proportion of infected sea fans on the entire reef is actually quite small, and the researchers just happened to find an infected area. To avoid this, the researchers should be careful to sample sites far enough apart to make contagion unlikely.

<sup>8</sup> When we are being very careful we say, "95% of samples of this size will produce confidence intervals that capture the true proportion of infected sea fans on the Las Redes Reef."



## CONNECTIONS

Now we can see a practical application of sampling distributions. To find a confidence interval, we lay out an interval measured in standard deviations. We're using the standard deviation as a ruler again. But now the standard deviation we need is the standard deviation of the sampling distribution. That's the one that tells how much the proportion varies. (And when we estimate it from the data, we call it a standard error.)



## WHAT HAVE WE LEARNED?

The first 10 chapters of the book explored graphical and numerical ways of summarizing and presenting sample data. We've learned (at last!) to use the sample we have at hand to say something about the *world at large*. This process, called statistical inference, is based on our understanding of sampling models and will be our focus for the rest of the book.

As our first step in statistical inference, we've learned to use our sample to make a *confidence interval* that estimates what proportion of a population has a certain characteristic.

We've learned that:

- ▶ Our best estimate of the true population proportion is the proportion we observed in the sample, so we center our confidence interval there.
- ▶ Samples don't represent the population perfectly, so we create our interval with a *margin of error*.
- ▶ This method successfully captures the true population proportion most of the time, providing us with a level of confidence in our interval.
- ▶ The higher the level of confidence we want, the *wider* our confidence interval becomes.
- ▶ The larger the sample size we have, the *narrower* our confidence interval can be.
- ▶ When designing a study, we can calculate the sample size we'll need to be able to reach conclusions that have a desired degree of precision and level of confidence.
- ▶ There are important assumptions and conditions we must check before using this (or any) statistical inference procedure.

We've learned to interpret a confidence interval by *Telling* what we believe is true in the entire population from which we took our random sample. Of course, we can't be *certain*. We've learned not to overstate or misinterpret what the confidence interval says.

## Terms

### Standard error

440. When we estimate the standard deviation of a sampling distribution using statistics found from the data, the estimate is called a standard error.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

### Confidence interval

441. A level C confidence interval for a model parameter is an interval of values usually of the form

$$\text{estimate} \pm \text{margin of error}$$

found from data in such a way that C% of all random samples will yield intervals that capture the true parameter value.

### One-proportion z-interval

442–444. A confidence interval for the true value of a proportion. The confidence interval is

$$\hat{p} \pm z^*SE(\hat{p}),$$

where  $z^*$  is a critical value from the Standard Normal model corresponding to the specified confidence level.

**Margin of error** 443. In a confidence interval, the extent of the interval on either side of the observed statistic value is called the margin of error. A margin of error is typically the product of a critical value from the sampling distribution and a standard error from the data. A small margin of error corresponds to a confidence interval that pins down the parameter precisely. A large margin of error corresponds to a confidence interval that gives relatively little information about the estimated parameter. For a proportion,

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**Critical value** 445. The number of standard errors to move away from the mean of the sampling distribution to correspond to the specified level of confidence. The critical value, denoted  $z^*$ , is usually found from a table or with technology.

## Skills

THINK

- ▶ Understand confidence intervals as a balance between the precision and the certainty of a statement about a model parameter.
- ▶ Understand that the margin of error of a confidence interval for a proportion changes with the sample size and the level of confidence.
- ▶ Know how to examine your data for violations of conditions that would make inference about a population proportion unwise or invalid.

SHOW

- ▶ Be able to construct a one-proportion  $z$ -interval.

TELL

- ▶ Be able to interpret a one-proportion  $z$ -interval in a simple sentence or two. Write such an interpretation so that it does not state or suggest that the parameter of interest is itself random, but rather that the bounds of the confidence interval are the random quantities about which we state our degree of confidence.

## CONFIDENCE INTERVALS FOR PROPORTIONS ON THE COMPUTER

Confidence intervals for proportions are so easy and natural that many statistics packages don't offer special commands for them. Most statistics programs want the "raw data" for computations. For proportions, the raw data are the "success" and "failure" status for each case. Usually, these are given as 1 or 0, but they might be category names like "yes" and "no." Often we just know the proportion of successes,  $\hat{p}$ , and the total count,  $n$ . Computer packages don't usually deal with summary data like this easily, but the statistics routines found on many graphing calculators allow you to create confidence intervals from summaries of the data—usually all you need to enter are the number of successes and the sample size.

In some programs you can reconstruct variables of 0's and 1's with the given proportions. But even when you have (or can reconstruct) the raw data values, you may not get exactly the same margin of error from a computer package as you would find working by hand. The reason is that some packages make approximations or use other methods. The result is very close but not exactly the same. Fortunately, Statistics means never having to say you're certain, so the approximate result is good enough.

## EXERCISES

- Margin of error.** A TV newscaster reports the results of a poll of voters, and then says, "The margin of error is plus or minus 4%." Explain carefully what that means.
- Margin of error.** A medical researcher estimates the percentage of children exposed to lead-base paint, adding that he believes his estimate has a margin of error of about 3%. Explain what the margin of error means.
- Conditions.** For each situation described below, identify the population and the sample, explain what  $p$  and  $\hat{p}$  represent, and tell whether the methods of this chapter can be used to create a confidence interval.
  - Police set up an auto checkpoint at which drivers are stopped and their cars inspected for safety problems. They find that 14 of the 134 cars stopped have at least one safety violation. They want to estimate the percentage of all cars that may be unsafe.
  - A TV talk show asks viewers to register their opinions on prayer in schools by logging on to a Web site. Of the 602 people who voted, 488 favored prayer in schools. We want to estimate the level of support among the general public.
  - A school is considering requiring students to wear uniforms. The PTA surveys parent opinion by sending a questionnaire home with all 1245 students; 380 surveys are returned, with 228 families in favor of the change.
  - A college admits 1632 freshmen one year, and four years later 1388 of them graduate on time. The college wants to estimate the percentage of all their freshman enrollees who graduate on time.
- More conditions.** Consider each situation described. Identify the population and the sample, explain what  $p$  and  $\hat{p}$  represent, and tell whether the methods of this chapter can be used to create a confidence interval.
  - A consumer group hoping to assess customer experiences with auto dealers surveys 167 people who recently bought new cars; 3% of them expressed dissatisfaction with the salesperson.
  - What percent of college students have cell phones? 2883 students were asked as they entered a football stadium, and 243 said they had phones with them.
  - 240 potato plants in a field in Maine are randomly checked, and only 7 show signs of blight. How severe is the blight problem for the U.S. potato industry?
  - 12 of the 309 employees of a small company suffered an injury on the job last year. What can the company expect in future years?
- Conclusions.** A catalog sales company promises to deliver orders placed on the Internet within 3 days. Follow-up calls to a few randomly selected customers show that a 95% confidence interval for the proportion of all orders that arrive on time is  $88\% \pm 6\%$ . What does this mean? Are these conclusions correct? Explain.
  - Between 82% and 94% of all orders arrive on time.
  - 95% of all random samples of customers will show that 88% of orders arrive on time.
  - 95% of all random samples of customers will show that 82% to 94% of orders arrive on time.
  - We are 95% sure that between 82% and 94% of the orders placed by the sampled customers arrived on time.
  - On 95% of the days, between 82% and 94% of the orders will arrive on time.
- More conclusions.** In January 2002, two students made worldwide headlines by spinning a Belgian euro 250 times and getting 140 heads—that's 56%. That makes the 90% confidence interval (51%, 61%). What does this mean? Are these conclusions correct? Explain.
  - Between 51% and 61% of all euros are unfair.
  - We are 90% sure that in this experiment this euro landed heads on between 51% and 61% of the spins.
  - We are 90% sure that spun euros will land heads between 51% and 61% of the time.
  - If you spin a euro many times, you can be 90% sure of getting between 51% and 61% heads.
  - 90% of all spun euros will land heads between 51% and 61% of the time.
- Confidence intervals.** Several factors are involved in the creation of a confidence interval. Among them are the sample size, the level of confidence, and the margin of error. Which statements are true?
  - For a given sample size, higher confidence means a smaller margin of error.
  - For a specified confidence level, larger samples provide smaller margins of error.
  - For a fixed margin of error, larger samples provide greater confidence.
  - For a given confidence level, halving the margin of error requires a sample twice as large.
- Confidence intervals, again.** Several factors are involved in the creation of a confidence interval. Among them are the sample size, the level of confidence, and the margin of error. Which statements are true?
  - For a given sample size, reducing the margin of error will mean lower confidence.
  - For a certain confidence level, you can get a smaller margin of error by selecting a bigger sample.
  - For a fixed margin of error, smaller samples will mean lower confidence.
  - For a given confidence level, a sample 9 times as large will make a margin of error one third as big.
- Cars.** What fraction of cars is made in Japan? The computer output below summarizes the results of a random sample of 50 autos. Explain carefully what it tells you.
 

z-Inter val for propor tion  
With 90.00% confidence,  
0.29938661 < p[japan] < 0.46984416

10. **Parole.** A study of 902 decisions made by the Nebraska Board of Parole produced the following computer output. Assuming these cases are representative of all cases that may come before the Board, what can you conclude?
- z-Interval for proportion  
With 95.00% confidence,  
0.56100658 < p(parole) < 0.62524619
11. **Contaminated chicken.** In January 2007 *Consumer Reports* published their study of bacterial contamination of chicken sold in the United States. They purchased 525 broiler chickens from various kinds of food stores in 23 states and tested them for types of bacteria that cause food-borne illnesses. Laboratory results indicated that 83% of these chickens were infected with *Campylobacter*.
- Construct a 95% confidence interval.
  - Explain what your confidence interval says about chicken sold in the United States.
  - A spokesperson for the U.S. Department of Agriculture dismissed the *Consumer Reports* finding, saying, "That's 500 samples out of 9 billion chickens slaughtered a year. . . . With the small numbers they [tested], I don't know that one would want to change one's buying habits." Is this criticism valid? Explain.
12. **Contaminated chicken, second course.** The January 2007 *Consumer Reports* study described in Exercise 11 also found that 15% of the 525 broiler chickens tested were infected with *Salmonella*.
- Are the conditions for creating a confidence interval satisfied? Explain.
  - Construct a 95% confidence interval.
  - Explain what your confidence interval says about chicken sold in the United States.
13. **Baseball fans.** In a poll taken in March of 2007, Gallup asked 1006 national adults whether they were baseball fans. 36% said they were. A year previously, 37% of a similar-size sample had reported being baseball fans.
- Find the margin of error for the 2007 poll if we want 90% confidence in our estimate of the percent of national adults who are baseball fans.
  - Explain what that margin of error means.
  - If we wanted to be 99% confident, would the margin of error be larger or smaller? Explain.
  - Find that margin of error.
  - In general, if all other aspects of the situation remain the same, will smaller margins of error produce greater or less confidence in the interval?
  - Do you think there's been a change from 2006 to 2007 in the real proportion of national adults who are baseball fans? Explain.
14. **Cloning 2007.** A May 2007 Gallup Poll found that only 11% of a random sample of 1003 adults approved of attempts to clone a human.
- Find the margin of error for this poll if we want 95% confidence in our estimate of the percent of American adults who approve of cloning humans.
  - Explain what that margin of error means.
  - If we only need to be 90% confident, will the margin of error be larger or smaller? Explain.
  - Find that margin of error.
- e) In general, if all other aspects of the situation remain the same, would smaller samples produce smaller or larger margins of error?
15. **Contributions, please.** The Paralyzed Veterans of America is a philanthropic organization that relies on contributions. They send free mailing labels and greeting cards to potential donors on their list and ask for a voluntary contribution. To test a new campaign, they recently sent letters to a random sample of 100,000 potential donors and received 4781 donations.
- Give a 95% confidence interval for the true proportion of their entire mailing list who may donate.
  - A staff member thinks that the true rate is 5%. Given the confidence interval you found, do you find that percentage plausible?
16. **Take the offer.** First USA, a major credit card company, is planning a new offer for their current cardholders. The offer will give double airline miles on purchases for the next 6 months if the cardholder goes online and registers for the offer. To test the effectiveness of the campaign, First USA recently sent out offers to a random sample of 50,000 cardholders. Of those, 1184 registered.
- Give a 95% confidence interval for the true proportion of those cardholders who will register for the offer.
  - If the acceptance rate is only 2% or less, the campaign won't be worth the expense. Given the confidence interval you found, what would you say?
17. **Teenage drivers.** An insurance company checks police records on 582 accidents selected at random and notes that teenagers were at the wheel in 91 of them.
- Create a 95% confidence interval for the percentage of all auto accidents that involve teenage drivers.
  - Explain what your interval means.
  - Explain what "95% confidence" means.
  - A politician urging tighter restrictions on drivers' licenses issued to teens says, "In one of every five auto accidents, a teenager is behind the wheel." Does your confidence interval support or contradict this statement? Explain.
18. **Junk mail.** Direct mail advertisers send solicitations (a.k.a. "junk mail") to thousands of potential customers in the hope that some will buy the company's product. The acceptance rate is usually quite low. Suppose a company wants to test the response to a new flyer, and sends it to 1000 people randomly selected from their mailing list of over 200,000 people. They get orders from 123 of the recipients.
- Create a 90% confidence interval for the percentage of people the company contacts who may buy something.
  - Explain what this interval means.
  - Explain what "90% confidence" means.
  - The company must decide whether to now do a mass mailing. The mailing won't be cost-effective unless it produces at least a 5% return. What does your confidence interval suggest? Explain.
19. **Safe food.** Some food retailers propose subjecting food to a low level of radiation in order to improve safety, but sale of such "irradiated" food is opposed by many people. Suppose a grocer wants to find out what his customers think. He has cashiers distribute surveys at checkout and

ask customers to fill them out and drop them in a box near the front door. He gets responses from 122 customers, of whom 78 oppose the radiation treatments. What can the grocer conclude about the opinions of all his customers?

20. **Local news.** The mayor of a small city has suggested that the state locate a new prison there, arguing that the construction project and resulting jobs will be good for the local economy. A total of 183 residents show up for a public hearing on the proposal, and a show of hands finds only 31 in favor of the prison project. What can the city council conclude about public support for the mayor's initiative?
21. **Death penalty, again.** In the survey on the death penalty you read about in the chapter, the Gallup Poll actually split the sample at random, asking 510 respondents the question quoted earlier, "Generally speaking, do you believe the death penalty is applied fairly or unfairly in this country today?" The other 510 were asked "Generally speaking, do you believe the death penalty is applied unfairly or fairly in this country today?" Seems like the same question, but sometimes the order of the choices matters. Suppose that for the second way of phrasing it, only 54% said they thought the death penalty was fairly applied.
- What kind of bias may be present here?
  - If we combine them, considering the overall group to be one larger random sample of 1020 respondents, what is a 95% confidence interval for the proportion of the general public that thinks the death penalty is being fairly applied?
  - How does the margin of error based on this pooled sample compare with the margins of error from the separate groups? Why?
22. **Gambling.** A city ballot includes a local initiative that would legalize gambling. The issue is hotly contested, and two groups decide to conduct polls to predict the outcome. The local newspaper finds that 53% of 1200 randomly selected voters plan to vote "yes," while a college Statistics class finds 54% of 450 randomly selected voters in support. Both groups will create 95% confidence intervals.
- Without finding the confidence intervals, explain which one will have the larger margin of error.
  - Find both confidence intervals.
  - Which group concludes that the outcome is too close to call? Why?
23. **Rickets.** Vitamin D, whether ingested as a dietary supplement or produced naturally when sunlight falls on the skin, is essential for strong, healthy bones. The bone disease rickets was largely eliminated in England during the 1950s, but now there is concern that a generation of children more likely to watch TV or play computer games than spend time outdoors is at increased risk. A recent study of 2700 children randomly selected from all parts of England found 20% of them deficient in vitamin D.
- Find a 98% confidence interval.
  - Explain carefully what your interval means.
  - Explain what "98% confidence" means.
24. **Pregnancy.** In 1998 a San Diego reproductive clinic reported 49 live births to 207 women under the age of 40 who had previously been unable to conceive.
- Find a 90% confidence interval for the success rate at this clinic.
  - Interpret your interval in this context.
  - Explain what "90% confidence" means.
  - Do these data refute the clinic's claim of a 25% success rate? Explain.
25. **Payments.** In a May 2007 Experian/Gallup Personal Credit Index poll of 1008 U.S. adults aged 18 and over, 8% of respondents said they were very uncomfortable with their ability to make their monthly payments on their current debt during the next three months. A more detailed poll surveyed 1288 adults, reporting similar overall results and also noting differences among four age groups: 18–29, 30–49, 50–64, and 65+.
- Do you expect the 95% confidence interval for the true proportion of all 18- to 29-year-olds who are worried to be wider or narrower than the 95% confidence interval for the true proportion of all U.S. consumers? Explain.
  - Do you expect this second poll's overall margin of error to be larger or smaller than the Experian/Gallup poll's? Explain.
26. **Back to campus again.** In 2004 ACT, Inc., reported that 74% of 1644 randomly selected college freshmen returned to college the next year. The study was stratified by type of college—public or private. The retention rates were 71.9% among 505 students enrolled in public colleges and 74.9% among 1139 students enrolled in private colleges.
- Will the 95% confidence interval for the true national retention rate in private colleges be wider or narrower than the 95% confidence interval for the retention rate in public colleges? Explain.
  - Do you expect the margin of error for the overall retention rate to be larger or smaller? Explain.
27. **Deer ticks.** Wildlife biologists inspect 153 deer taken by hunters and find 32 of them carrying ticks that test positive for Lyme disease.
- Create a 90% confidence interval for the percentage of deer that may carry such ticks.
  - If the scientists want to cut the margin of error in half, how many deer must they inspect?
  - What concerns do you have about this sample?
28. **Pregnancy, II.** The San Diego reproductive clinic in Exercise 24 wants to publish updated information on its success rate.
- The clinic wants to cut the stated margin of error in half. How many patients' results must be used?
  - Do you have any concerns about this sample? Explain.
29. **Graduation.** It's believed that as many as 25% of adults over 50 never graduated from high school. We wish to see if this percentage is the same among the 25 to 30 age group.
- How many of this younger age group must we survey in order to estimate the proportion of non-grads to within 6% with 90% confidence?
  - Suppose we want to cut the margin of error to 4%. What's the necessary sample size?
  - What sample size would produce a margin of error of 3%?



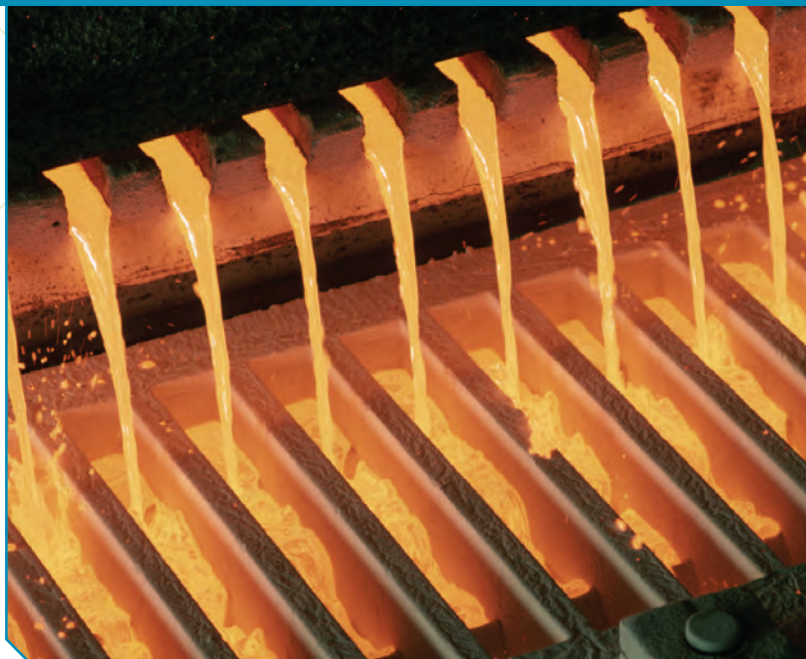
30. **Hiring.** In preparing a report on the economy, we need to estimate the percentage of businesses that plan to hire additional employees in the next 60 days.
- How many randomly selected employers must we contact in order to create an estimate in which we are 98% confident with a margin of error of 5%?
  - Suppose we want to reduce the margin of error to 3%. What sample size will suffice?
  - Why might it not be worth the effort to try to get an interval with a margin of error of only 1%?
31. **Graduation, again.** As in Exercise 29, we hope to estimate the percentage of adults aged 25 to 30 who never graduated from high school. What sample size would allow us to increase our confidence level to 95% while reducing the margin of error to only 2%?
32. **Better hiring info.** Editors of the business report in Exercise 30 are willing to accept a margin of error of 4% but want 99% confidence. How many randomly selected employers will they need to contact?
33. **Pilot study.** A state's environmental agency worries that many cars may be violating clean air emissions standards. The agency hopes to check a sample of vehicles in order to estimate that percentage with a margin of error of 3% and 90% confidence. To gauge the size of the problem, the agency first picks 60 cars and finds 9 with faulty emissions systems. How many should be sampled for a full investigation?
34. **Another pilot study.** During routine screening, a doctor notices that 22% of her adult patients show higher than normal levels of glucose in their blood—a possible warning signal for diabetes. Hearing this, some medical researchers decide to conduct a large-scale study, hoping to estimate the proportion to within 4% with 98% confidence. How many randomly selected adults must they test?
35. **Approval rating.** A newspaper reports that the governor's approval rating stands at 65%. The article adds that the poll is based on a random sample of 972 adults and has a margin of error of 2.5%. What level of confidence did the pollsters use?
36. **Amendment.** A TV news reporter says that a proposed constitutional amendment is likely to win approval in the upcoming election because a poll of 1505 likely voters indicated that 52% would vote in favor. The reporter goes on to say that the margin of error for this poll was 3%.
- Explain why the poll is actually inconclusive.
  - What confidence level did the pollsters use?



### JUST CHECKING Answers

- No. We know that in the sample 17% said "yes"; there's no need for a margin of error.
- No, we are 95% confident that the percentage falls in some interval, not exactly on a particular value.
- Yes. That's what the confidence interval means.
- No. We don't know for sure that's true; we are only 95% confident.
- No. That's our level of confidence, not the proportion of people receiving unsolicited text messages. The sample suggests the proportion is much lower.
- Wider.
- Lower.
- Smaller.

# Testing Hypotheses About Proportions



## AS

### Activity: Testing a Claim.

Can we really draw a reasonable conclusion from a random sample? Run this simulation before you read the chapter, and you'll gain a solid sense of what we're doing here.

Ingots are huge pieces of metal, often weighing more than 20,000 pounds, made in a giant mold. They must be cast in one large piece for use in fabricating large structural parts for cars and planes. If they crack while being made, the crack can propagate into the zone required for the part, compromising its integrity. Airplane manufacturers insist that metal for their planes be defect-free, so the ingot must be made over if any cracking is detected.

Even though the metal from the cracked ingot is recycled, the scrap cost runs into the tens of thousands of dollars. Metal manufacturers would like to avoid cracking if at all possible. But the casting process is complicated and not everything is completely under control. In one plant, only about 80% of the ingots have been free of cracks. In an attempt to reduce the cracking proportion, the plant engineers and chemists recently tried out some changes in the casting process. Since then, 400 ingots have been cast and only 17% of them have cracked. Should management declare victory? Has the cracking rate really decreased, or was 17% just due to luck?

We can treat the 400 ingots cast with the new method as a random sample. We know that each random sample will have a somewhat different proportion of cracked ingots. Is the 17% we observe merely a result of natural sampling variability, or is this lower cracking rate strong enough evidence to assure management that the true cracking rate now is really below 20%?

People want answers to questions like these all the time. Has the president's approval rating changed since last month? Has teenage smoking decreased in the past five years? Is the global temperature increasing? Did the Super Bowl ad we bought actually increase sales? To answer such questions, we test *hypotheses* about models.

*"Half the money I spend on advertising is wasted; the trouble is I don't know which half."*

—John Wanamaker  
(attributed)

## Hypotheses

How can we state and test a hypothesis about ingot cracking? Hypotheses are working models that we adopt temporarily. To test whether the changes made by the engineers have *improved* the cracking rate, we assume that they have in fact

Hypothesis *n.*;  
pl. {Hypotheses}.

A supposition; a proposition or principle which is supposed or taken for granted, in order to draw a conclusion or inference for proof of the point in question; something not proved, but assumed for the purpose of argument.  
—*Webster's Unabridged Dictionary, 1913*

### NOTATION ALERT:

Capital H is the standard letter for hypotheses.  $H_0$  always labels the null hypothesis, and  $H_A$  labels the alternative hypothesis.

To remind us that the parameter value comes from the null hypothesis, it is sometimes written as  $p_0$  and the standard deviation as

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$$

made no difference and that any apparent improvement is just random fluctuation (sampling error). So, our starting hypothesis, called the **null hypothesis**, is that the proportion of cracks is still 20%.

The null hypothesis, which we denote  $H_0$ , specifies a population model parameter of interest and proposes a value for that parameter. We usually write down the null hypothesis in the form  $H_0: \text{parameter} = \text{hypothesized value}$ . This is a concise way to specify the two things we need most: the identity of the parameter we hope to learn about and a specific hypothesized value for that parameter. (We need a hypothesized value so we can compare our observed statistic value to it.)

Which value to use is often obvious from the *Who* and *What* of the data. But sometimes it takes a bit of thinking to translate the question we hope to answer into a hypothesis about a parameter. For the ingots we can write  $H_0: p = 0.20$ .

The alternative hypothesis, which we denote  $H_A$ , contains the values of the parameter that we consider plausible if we reject the null hypothesis. In the ingots example, our null hypothesis is that  $p = 0.20$ . What's the alternative? Management is interested in *reducing* the cracking rate, so their alternative is  $H_A: p < 0.20$ .

What would convince you that the cracking rate had actually gone down? If you observed a cracking rate *much lower* than 20% in your sample, you'd likely be convinced. If only 3 out of the next 400 ingots crack (for a rate of 0.75%), most folks would conclude that the changes helped. But if the sample cracking rate is only slightly lower than 20%, you should be skeptical. After all, observed proportions do vary, so we wouldn't be surprised to see some difference. How much smaller must the cracking rate be before we *are* convinced that it has changed? Whenever we ask about the size of a statistical difference, we naturally think of using the standard deviation as a ruler. So let's start by finding the standard deviation of the sample cracking rate.

Since the company changed the process, 400 new ingots have been cast. The sample size of 400 is big enough to satisfy the **Success/Failure Condition**. (We expect  $0.20 \times 400 = 80$  ingots to crack.) We have no reason to think the ingots are not independent, so the Normal sampling distribution model should work well. The standard deviation of the sampling model is

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{400}} = 0.02$$

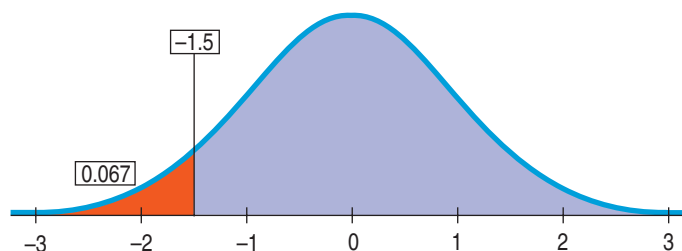
**Why is this a standard deviation and not a standard error?** Because we haven't estimated anything. When we assume that the null hypothesis is true, it gives us a value for the model parameter  $p$ . With proportions, if we know  $p$ , then we also automatically know its standard deviation. And because we find the standard deviation from the model parameter, this is a standard deviation and not a standard error. When we found a confidence interval for  $p$ , we could not assume that we knew its value, so we estimated the standard deviation from the sample value  $\hat{p}$ .

Now we know both parameters of the Normal sampling distribution model:  $p = 0.20$  and  $SD(\hat{p}) = 0.02$ , so we can find out how likely it would be to see the observed value of  $\hat{p} = 17\%$ . Since we are using a Normal model, we find the *z*-score:

$$z = \frac{0.17 - 0.20}{0.02} = -1.5$$

Then we ask, "How likely is it to observe a value at least 1.5 standard deviations below the mean of a Normal model?" The answer (from a calculator, computer program, or the Normal table) is about 0.067. This is the probability of observing a cracking rate of 17% or less in a sample of 400 if the null hypothesis is true.

Management now must decide whether an event that would happen 6.7% of the time by chance is strong enough evidence to conclude that the true cracking proportion has decreased.



**FIGURE 20.1**

How likely is a z-score of  $-1.5$  (or lower)? This is what it looks like. The red area is 0.067 of the total area under the curve.

## A Trial as a Hypothesis Test

Does the reasoning of hypothesis tests seem backward? That could be because we usually prefer to think about getting things right rather than getting them wrong. You have seen this reasoning before in a different context. This is the logic of jury trials.

Let's suppose a defendant has been accused of robbery. In British common law and those systems derived from it (including U.S. law), the null hypothesis is that the defendant is innocent. Instructions to juries are quite explicit about this.

The evidence takes the form of facts that seem to contradict the presumption of innocence. For us, this means collecting data. In the trial, the prosecutor presents evidence. ("If the defendant were innocent, wouldn't it be remarkable that the police found him at the scene of the crime with a bag full of money in his hand, a mask on his face, and a getaway car parked outside?")

The next step is to judge the evidence. Evaluating the evidence is the responsibility of the jury in a trial, but it falls on your shoulders in hypothesis testing. The jury considers the evidence in light of the *presumption* of innocence and judges whether the evidence against the defendant would be plausible *if the defendant were in fact innocent*.

Like the jury, you ask, "Could these data plausibly have happened by chance if the null hypothesis were true?" If they are very unlikely to have occurred, then the evidence raises a reasonable doubt about the null hypothesis.

Ultimately, you must make a decision. The standard of "beyond a reasonable doubt" is wonderfully ambiguous because it leaves the jury to decide the degree to which the evidence contradicts the hypothesis of innocence. Juries don't explicitly use probability to help them decide whether to reject that hypothesis. But when you ask the same question of your null hypothesis, you have the advantage of being able to quantify exactly how surprising the evidence would be were the null hypothesis true.

How unlikely is unlikely? Some people set rigid standards, like 1 time out of 20 (0.05) or 1 time out of 100 (0.01). But if *you* have to make the decision, you must judge for yourself in each situation whether the probability of observing your data is small enough to constitute "reasonable doubt."

**AS** **Activity: The Reasoning of Hypothesis Testing.** Our reasoning is based on a rule of logic that dates back to ancient scholars. Here's a modern discussion of it.

## P-Values

The fundamental step in our reasoning is the question "Are the data surprising, given the null hypothesis?" And the key calculation is to determine exactly how likely the data we observed would be were the null hypothesis a true model of the world. So we need a *probability*. Specifically, we want to find the probability of seeing data like these (or something even less likely) *given* that the null hypothesis is true. Statisticians are so thrilled with their ability to measure precisely

**Beyond a Reasonable Doubt**

We ask whether the data were unlikely beyond a reasonable doubt. We've just calculated that probability. The probability that the observed statistic value (or an even more extreme value) could occur if the null model were true—in this case, 0.067—is the P-value.

**NOTATION ALERT:**

We have many P's to keep straight. We use an uppercase P for probabilities, as in  $P(A)$ , and for the special probability we care about in hypothesis testing, the P-value.

We use lowercase  $p$  to denote our model's underlying proportion parameter and  $\hat{p}$  to denote our observed proportion statistic.

how surprised they are that they give this probability a special name. It's called a **P-value**.<sup>1</sup>

When the P-value is high, we haven't seen anything unlikely or surprising at all. Events that have a high probability of happening happen often. The data are thus consistent with the model from the null hypothesis, and we have no reason to reject the null hypothesis. But we realize that many other similar hypotheses could also account for the data we've seen, so *we haven't proven that the null hypothesis is true*. The most we can say

is that it doesn't appear to be false. Formally, we "fail to reject" the null hypothesis. That's a pretty weak conclusion, but it's all we're entitled to.

When the P-value is low enough, it says that it's very unlikely we'd observe data like these if our null hypothesis were true. We started with a model. Now that model tells us that the data we have are unlikely to have happened. The model and data are at odds with each other, so we have to make a choice. Either the null hypothesis is correct and we've just seen something remarkable, or the null hypothesis is wrong, and we were wrong to use it as the basis for computing our P-value. Perhaps another model is correct, and the data really aren't that remarkable after all. If you believe in data more than in assumptions, then, given that choice, you should reject the null hypothesis.

## What to Do with an "Innocent" Defendant

*"If the People fail to satisfy their burden of proof, you must find the defendant not guilty."*

—NY state jury instructions

**Don't "Accept" the Null Hypothesis**

Every child knows that he (or she) is at the "center of the universe," so it's natural to suppose that the sun revolves around the earth. The fact that the sun appears to rise in the east every morning and set in the west every evening is *consistent* with this hypothesis and *seems* to lend support to it, but it certainly doesn't prove it, as we all eventually come to understand.

If the evidence is not strong enough to reject the defendant's presumption of innocence, what verdict does the jury return? They say "not guilty." Notice that they do not say that the defendant is innocent. All they say is that they have not seen sufficient evidence to convict, to reject innocence. The defendant may, in fact, be innocent, but the jury has no way to be sure.

Said statistically, the jury's null hypothesis is  $H_0$ : innocent defendant. If the evidence is too unlikely given this assumption, the jury rejects the null hypothesis and finds the defendant guilty. But—and this is an important distinction—if there is *insufficient evidence* to convict the defendant, the jury does not decide that  $H_0$  is true and declare the defendant innocent. Juries can only *fail to reject* the null hypothesis and declare the defendant "not guilty."

In the same way, if the data are not particularly unlikely under the assumption that the null hypothesis is true, then the most we can do is to "fail to reject" our null hypothesis. We never declare the null hypothesis to be true (or "accept" the null), because we simply do not know whether it's true or not. (After all, more evidence may come along later.)

In the trial, the burden of proof is on the prosecution. In a hypothesis test, the burden of proof is on the unusual claim. The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual and for which we must marshal evidence.

Imagine a clinical trial testing the effectiveness of a new headache remedy. In Chapter 13 we saw the value of comparing such treatments to a placebo. The null hypothesis, then, is that the new treatment is no more effective than the placebo. This is important, because some patients will improve even when administered the placebo treatment. If we use only six people to test the drug, the results are likely *not to be clear* and we'll be unable to reject the hypothesis. Does this mean the drug doesn't work? Of course not. It simply means that we don't have enough

<sup>1</sup> You'd think if they were so excited, they'd give it a better name, but "P-value" is about as excited as statisticians get.

evidence to reject our assumption. That's why we don't start by assuming that the drug *is more effective*. If we were to do that, then we could test just a few people, find that the results aren't clear, and claim that since we've been unable to reject our original assumption the drug must be effective. The FDA is unlikely to be impressed by that argument.



## JUST CHECKING

1. A research team wants to know if aspirin helps to thin blood. The null hypothesis says that it doesn't. They test 12 patients, observe the proportion with thinner blood, and get a P-value of 0.32. They proclaim that aspirin doesn't work. What would you say?
2. An allergy drug has been tested and found to give relief to 75% of the patients in a large clinical trial. Now the scientists want to see if the new, improved version works even better. What would the null hypothesis be?
3. The new drug is tested and the P-value is 0.0001. What would you conclude about the new drug?

## The Reasoning of Hypothesis Testing

*"The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."*

—Sir Ronald Fisher, *The Design of Experiments*

Some folks pronounce the hypothesis labels "Ho!" and "Ha!" (but it makes them seem overexcitable). We prefer to pronounce  $H_0$  "H naught" (as in "all is for naught").

Hypothesis tests follow a carefully structured path. To avoid getting lost as we navigate down it, we divide that path into four distinct sections.

### 1. HYPOTHESES

First we state the null hypothesis. That's usually the skeptical claim that nothing's different. Are we considering a (New! Improved!) possibly better method? The null hypothesis says, "Oh yeah? Convince me!" To convert a skeptic, we must pile up enough evidence against the null hypothesis that we can reasonably reject it.

In statistical hypothesis testing, hypotheses are almost always about model parameters. To assess how unlikely our data may be, we need a null model. The null hypothesis specifies a particular parameter value to use in our model. In the usual shorthand, we write  $H_0$ : *parameter = hypothesized value*. The **alternative hypothesis**,  $H_A$ , contains the values of the parameter we consider plausible when we reject the null.

### FOR EXAMPLE

#### Writing hypotheses

A large city's Department of Motor Vehicles claimed that 80% of candidates pass driving tests, but a newspaper reporter's survey of 90 randomly selected local teens who had taken the test found only 61 who passed.

**Question:** Does this finding suggest that the passing rate for teenagers is lower than the DMV reported? Write appropriate hypotheses.

I'll assume that the passing rate for teenagers is the same as the DMV's overall rate of 80%, unless there's strong evidence that it's lower.

$$H_0: p = 0.80$$

$$H_A: p < 0.80$$

## 2. MODEL

To plan a statistical hypothesis test, specify the *model* you will use to test the null hypothesis and the parameter of interest. Of course, all models require assumptions, so you will need to state them and check any corresponding conditions.

Your Model step should end with a statement such as

*Because the conditions are satisfied, I can model the sampling distribution of the proportion with a Normal model.*

Watch out, though. Your Model step could end with

*Because the conditions are not satisfied, I can't proceed with the test. (If that's the case, stop and reconsider.)*

Each test in the book has a name that you should include in your report. We'll see many tests in the chapters that follow. Some will be about more than one sample, some will involve statistics other than proportions, and some will use models other than the Normal (and so will not use z-scores). **The test about proportions is called a one-proportion z-test.<sup>2</sup>**

### When the Conditions Fail . . .

You might proceed with caution, explicitly stating your concerns. Or you may need to do the analysis with and without an outlier, or on different subgroups, or after re-expressing the response variable. Or you may not be able to proceed at all.

### AS Activity: Was the Observed Outcome Unlikely?

Complete the test you started in the first activity for this chapter. The narration explains the steps of the hypothesis test.

### ONE-PROPORTION z-TEST

The conditions for the one-proportion z-test are the same as for the one-proportion z-interval. We test the hypothesis  $H_0: p = p_0$  using the statistic  $z = \frac{(\hat{p} - p_0)}{SD(\hat{p})}$ . We use the hypothesized proportion to find the

standard deviation,  $SD(\hat{p}) = \sqrt{\frac{p_0q_0}{n}}$ .

When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

## FOR EXAMPLE

### Checking the conditions

**Recap:** A large city's DMV claimed that 80% of candidates pass driving tests. A reporter has results from a survey of 90 randomly selected local teens who had taken the test.

**Question:** Are the conditions for inference satisfied?

- ✓ The 90 teens surveyed were a random sample of local teenage driving candidates.
- ✓ 90 is fewer than 10% of the teenagers who take driving tests in a large city.
- ✓ We expect  $np_0 = 90(0.80) = 72$  successes and  $nq_0 = 90(0.20) = 18$  failures. Both are at least 10.

The conditions are satisfied, so it's okay to use a Normal model and perform a one-proportion z-test.

### Conditional Probability

Did you notice that a P-value is a conditional probability? It's the probability that the observed results could have happened *if the null hypothesis is true*.

## 3. MECHANICS

Under "Mechanics," we place the actual calculation of our test statistic from the data. Different tests we encounter will have different formulas and different test statistics. Usually, the mechanics are handled by a statistics program or calculator, but it's good to have the formulas recorded for reference and to know what's

<sup>2</sup> It's also called the "one-sample test for a proportion."

being computed. The ultimate goal of the calculation is to obtain a P-value—the probability that the observed statistic value (or an even more extreme value) occur if the null model is correct. If the P-value is small enough, we'll reject the null hypothesis.

## FOR EXAMPLE

### Finding a P-value

**Recap:** A large city's DMV claimed that 80% of candidates pass driving tests, but a survey of 90 randomly selected local teens who had taken the test found only 61 who passed.

**Question:** What's the P-value for the one-proportion z-test?

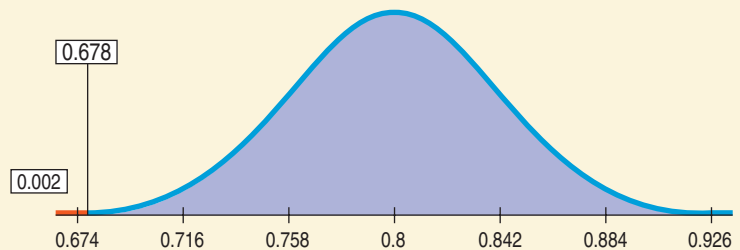
I have  $n = 90$ ,  $x = 61$ , and a hypothesized  $p = 0.80$ .

$$\hat{p} = \frac{61}{90} \approx 0.678$$

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.8)(0.2)}{90}} \approx 0.042$$

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.678 - 0.800}{0.042} \approx -2.90$$

$$P\text{-value} = P(z < -2.90) = 0.002$$



## 4. CONCLUSION

The conclusion in a hypothesis test is always a statement about the null hypothesis. The conclusion must state either that we reject or that we fail to reject the null hypothesis. And, as always, the conclusion should be stated in context.

## FOR EXAMPLE

### Stating the conclusion

**Recap:** A large city's DMV claimed that 80% of candidates pass driving tests. Data from a reporter's survey of randomly selected local teens who had taken the test produced a P-value of 0.002.

**Question:** What can the reporter conclude? And how might the reporter explain what the P-value means for the newspaper story?

Because the P-value of 0.002 is very low, I reject the null hypothesis. These survey data provide strong evidence that the passing rate for teenagers taking the driving test is lower than 80%.

If the passing rate for teenage driving candidates were actually 80%, we'd expect to see success rates this low in only about 1 in 500 samples (0.2%). This seems quite unlikely, casting doubt that the DMV's stated success rate applies to teens.

“. . . They make things  
admirably plain,  
But one hard question will  
remain:  
If one hypothesis you lose,  
Another in its place you  
choose . . .”

—James Russell Lowell,  
*Credidimus Jovem  
Regnare*

Your conclusion about the null hypothesis should never be the end of a testing procedure. Often there are actions to take or policies to change. In our ingot example, management must decide whether to continue the changes proposed by the engineers. The decision always includes the practical consideration of whether the new method is worth the cost. Suppose management decides to reject the null hypothesis of 20% cracking in favor of the alternative that the percentage has been reduced. They must still evaluate how much the cracking rate has been reduced and how much it cost to accomplish the reduction. The *size of the effect* is always a concern when we test hypotheses. A good way to look at the effect size is to examine a confidence interval.



**How much does it cost?** Formal tests of a null hypothesis base the decision of whether to reject the null hypothesis solely on the size of the P-value. But in real life, we want to evaluate the costs of our decisions as well. How much would you be willing to pay for a faster computer? Shouldn't your decision depend on how much faster? And on how much more it costs? Costs are not just monetary either. Would you use the same standard of proof for testing the safety of an airplane as for the speed of your new computer?

## Alternative Alternatives

Tests on the ingot data can be viewed in two different ways. We know the old cracking rate is 20%, so the null hypothesis is

$$H_0: p = 0.20$$

**AS** **Activity: the Alternative Hypotheses.** This interactive tool provides easy ways to visualize how one- and two-tailed alternative hypotheses work.

But we have a choice of alternative hypotheses. A metallurgist working for the company might be interested in *any* change in the cracking rate due to the new process. Even if the rate got worse, she might learn something useful from it. She's interested in possible changes on both sides of the null hypothesis. So she would write her alternative hypothesis as

$$H_A: p \neq 0.20$$

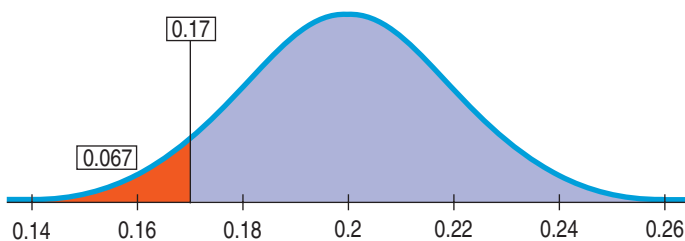
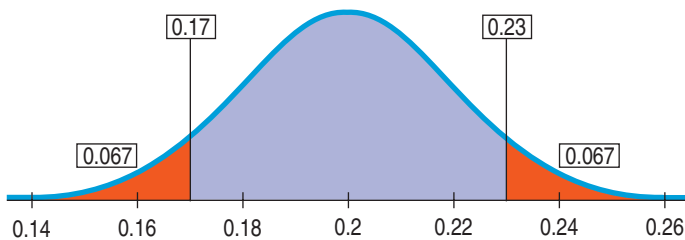
An alternative hypothesis such as this is known as a **two-sided alternative**,<sup>3</sup> because we are equally interested in deviations on either side of the null hypothesis value. For two-sided alternatives, the P-value is the probability of deviating in *either* direction from the null hypothesis value.

But management is really interested only in *lowering* the cracking rate below 20%. The scientific value of knowing how to *increase* the cracking rate may not appeal to them. The only alternative of interest to them is that the cracking rate *decreases*. They would write their alternative hypothesis as

$$H_A: p < 0.20$$

An alternative hypothesis that focuses on deviations from the null hypothesis value in only one direction is called a **one-sided alternative**.

For a hypothesis test with a one-sided alternative, the P-value is the probability of deviating *only in the direction of the alternative* away from the null hypothesis value. For the same data, the one-sided P-value is half the two-sided P-value. So, a one-sided test will reject the null hypothesis more often. If you aren't sure which to use, a two-sided test is always more conservative. Be sure you can justify the choice of a one-sided test from the *Why* of the situation.



<sup>3</sup> It is also called a **two-tailed alternative**, because the probabilities we care about are found in both tails of the sampling distribution.

## STEP-BY-STEP EXAMPLE

## Testing a Hypothesis

Anyone who plays or watches sports has heard of the “home field advantage.” Teams tend to win more often when they play at home. Or do they?

If there were no home field advantage, the home teams would win about half of all games played. In the 2007 Major League Baseball season, there were 2431 regular-season games. (Tied at the end of the regular season, the Colorado Rockies and San Diego Padres played an extra game to determine who won the Wild Card playoff spot.) It turns out that the home team won 1319 of the 2431 games, or 54.26% of the time.

**Question:** Could this deviation from 50% be explained just from natural sampling variability, or is it evidence to suggest that there really is a home field advantage, at least in professional baseball?



**Plan** State what we want to know.

Define the variables and discuss the W’s.

**Hypotheses** The null hypothesis makes the claim of no difference from the baseline. Here, that means no home field advantage.

We are interested only in a home field *advantage*, so the alternative hypothesis is one-sided.

**Model** Think about the assumptions and check the appropriate conditions.

I want to know whether the home team in professional baseball is more likely to win. The data are all 2431 games from the 2007 Major League Baseball season. The variable is whether or not the home team won. The parameter of interest is the proportion of home team wins. If there’s no advantage, I’d expect that proportion to be 0.50.

$$H_0: p = 0.50$$

$$H_A: p > 0.50$$

- ✓ **Independence Assumption:** Generally, the outcome of one game has no effect on the outcome of another game. But this may not be strictly true. For example, if a key player is injured, the probability that the team will win in the next couple of games may decrease slightly, but independence is still roughly true. The data come from one entire season, but I expect other seasons to be similar.
- ✓ **Randomization Condition:** I have results for all 2431 games of the 2007 season. But I’m not just interested in 2007, and those games, while not randomly selected, should be a reasonable representative sample of all Major League Baseball games in the recent past and near future.
- ✓ **10% Condition:** We are interested in home field advantage for Major League Baseball for all seasons. While not a random sample, these 2431 games are fewer than 10% of all games played over the years.
- ✓ **Success/Failure Condition:** Both  $np_0 = 2431(0.50) = 1215.5$  and  $nq_0 = 2431(0.50) = 1215.5$  are at least 10.

**AS**

**Activity: Practice with Testing Hypotheses About Proportions.** Here’s an interactive tool that makes it easy to see what’s going on in a hypothesis test.

Specify the sampling distribution model.

State what test you plan to use.

Because the conditions are satisfied, I'll use a Normal model for the sampling distribution of the proportion and do a **one-proportion z-test**.

SHOW

**Mechanics** The null model gives us the mean, and (because we are working with proportions) the mean gives us the standard deviation.

Next, we find the z-score for the observed proportion, to find out how many standard deviations it is from the hypothesized proportion.

From the z-score, we can find the P-value, which tells us the probability of observing a value that extreme (or more).

The probability of observing a value 4.20 or more standard deviations above the mean of a Normal model can be found by computer, calculator, or table to be  $< 0.001$ .

The null model is a Normal distribution with a mean of 0.50 and a standard deviation of

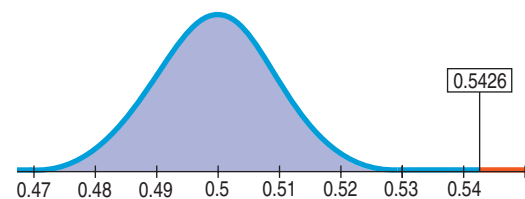
$$SD(\hat{p}) = \sqrt{\frac{p_0q_0}{n}} = \sqrt{\frac{(0.5)(1-0.5)}{2431}} = 0.01014$$

The observed proportion,  $\hat{p}$ , is 0.5426.

So the z-value is

$$z = \frac{0.5426 - 0.5}{0.01014} = 4.20$$

The sample proportion lies 4.20 standard deviations above the mean.



The corresponding P-value is  $< 0.001$ .

TELL

**Conclusion** State your conclusion about the parameter—in context, of course!

The P-value of  $< 0.001$  says that if the true proportion of home team wins were 0.50, then an observed value of 0.5426 (or larger) would occur less than 1 time in 1000. With a P-value so small, I reject  $H_0$ . I have evidence that the true proportion of home team wins is not 50%. It appears there is a home field advantage.

Ok, but how *big* is the home field advantage? Measuring the size of the effect involves a confidence interval. (Use your calculator.)

### TI Tips

## Testing a hypothesis

By now probably nothing surprises you about your calculator. Of course it can help you with the mechanics of a hypothesis test. But that's not much. It cannot write the correct hypotheses, check the appropriate conditions, interpret the results, or state a conclusion. You have to do the tough stuff!

```

EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:Interval...

```

```

1-PropZTest
P0:.5
x:1319
n:2431
PROP#P0 <P0 [X]P0
Calculate Draw

```

```

1-PropZTest
PROP>.5
z=4.198342507
P=1.3452178E-5
P̂=.542575072
n=2431

```

```

1-PropZInt
(.52277, .56238)
P̂=.542575072
n=2431

```

Let's do the mechanics of the Step-By-Step example about home field advantage in baseball. We hypothesized that home teams would win 50% of all games, but during this 2431-game season they actually won 54.26% of the time.

- Go to the **STAT TESTS** menu. Scroll down the list and select **5:1-Prop ZTest**.
- Specify the hypothesized proportion **P0**.
- Enter  $x$ , the observed number of wins: **1319**.
- Specify the sample size.
- Since this is a one-tail upper tail test, indicate that you want to see if the observed proportion is significantly greater than what was hypothesized.
- **Calculate** the result.

Ok, the rest is up to you. The calculator reports a z-score of 4.20 and a P-value (in scientific notation) of  $1.35 \times 10^{-5}$ , or about 0.00001. Such a small P-value indicates that the high percentage of home team wins is highly unlikely to be sampling error. State your conclusion in the appropriate context.

And how big is the advantage for the home team? In the last chapter you learned to create a 95% confidence interval. Try it here.

Looks like we can be 95% confident that in major league baseball games the home team wins between 52.3% and 56.2% of the time. Over a full season, the low end of this interval, 52.3% of the 81 home games, is nearly 2 extra victories, on average. The upper end, 56.2%, is 5 extra wins.

## P-Values and Decisions: What to Tell About a Hypothesis Test



Hypothesis tests are particularly useful when we must make a decision. Is the defendant guilty or not? Should we choose print advertising or television? Questions like these cannot always be answered with the margins of error of confidence intervals. The absolute nature of the hypothesis test decision, however, makes some people (including the authors) uneasy. If possible, it's often a good idea to report a confidence interval for the parameter of interest as well.

How small should the P-value be in order for you to reject the null hypothesis? A jury needs enough evidence to show the defendant guilty "beyond a reasonable doubt." How does that translate to P-values? The answer is that it's highly context-dependent. When we're screening for a disease and want to be sure we treat all those who are sick, we may be willing to reject the null hypothesis of no disease with a P-value as large as 0.10. We would rather treat the occasional healthy person than fail to treat someone who was really sick. But a long-standing hypothesis, believed by many to be true, needs stronger evidence (and a correspondingly small P-value) to reject it.

See if you require the same P-value to reject each of the following null hypotheses:

- ▶ A renowned musicologist claims that she can distinguish between the works of Mozart and Haydn simply by hearing a randomly selected 20 seconds of music from any work by either composer. What's the null hypothesis? If she's just guessing, she'll get 50% of the pieces correct, on average. So our null hypothesis is that  $p$  is 50%. If she's for real, she'll get more than 50% correct. Now, we present her with 10 pieces of Mozart or Haydn chosen at random. She gets 9 out of 10 correct. It turns out that the P-value associated with

“Extraordinary claims require extraordinary proof.”

—Carl Sagan

that result is 0.011. (In other words, if you tried to just guess, you’d get at least 9 out of 10 correct only about 1% of the time.) What would *you* conclude? Most people would probably reject the null hypothesis and be convinced that she has some ability to do as she claims. Why? Because the P-value is small and we don’t have any particular reason to doubt the alternative.

- ▶ On the other hand, imagine a student who bets that he can make a flipped coin land the way he wants just by thinking hard. To test him, we flip a fair coin 10 times. Suppose he gets 9 out of 10 right. This also has a P-value of 0.011. Are you willing now to reject this null hypothesis? Are you convinced that he’s not just lucky? What amount of evidence *would* convince you? We require more evidence if rejecting the null hypothesis would contradict long-standing beliefs or other scientific results. Of course, with sufficient evidence we would revise our opinions (and scientific theories). That’s how science makes progress.

Another factor in choosing a P-value is the importance of the issue being tested. Consider the following two tests:

- ▶ A researcher claims that the proportion of college students who hold part-time jobs now is higher than the proportion known to hold such jobs a decade ago. You might be willing to believe the claim (and reject the null hypothesis of no change) with a P-value of 10%.
- ▶ An engineer claims that the proportion of rivets holding the wing on an airplane that are likely to fail is below the proportion at which the wing would fall off. What P-value would be small enough to get you to fly on that plane?

**A S** **Activity: Hypothesis Tests for Proportions.** You’ve probably noticed that the tools for confidence intervals and for hypothesis tests are similar. See how tests and intervals for proportions are related—and an important way in which they differ.

Your conclusion about any null hypothesis should be accompanied by the P-value of the test. Don’t just declare the null hypothesis rejected or not rejected. Report the P-value to show the strength of the evidence against the hypothesis and the effect size. This will let each reader decide whether or not to reject the null hypothesis and whether or not to consider the result important if it is statistically significant.

To complete your analysis, follow your test with a confidence interval for the parameter of interest, to report the size of the effect.



## JUST CHECKING

4. A bank is testing a new method for getting delinquent customers to pay their past-due credit card bills. The standard way was to send a letter (costing about \$0.40) asking the customer to pay. That worked 30% of the time. They want to test a new method that involves sending a DVD to customers encouraging them to contact the bank and set up a payment plan. Developing and sending the video costs about \$10.00 per customer. What is the parameter of interest? What are the null and alternative hypotheses?
5. The bank sets up an experiment to test the effectiveness of the DVD. They mail it out to several randomly selected delinquent customers and keep track of how many actually do contact the bank to arrange payments. The bank’s statistician calculates a P-value of 0.003. What does this P-value suggest about the DVD?
6. The statistician tells the bank’s management that the results are clear and that they should switch to the DVD method. Do you agree? What else might you want to know?

## STEP-BY-STEP EXAMPLE

## Tests and Intervals

Advances in medical care such as prenatal ultrasound examination now make it possible to determine a child's sex early in a pregnancy. There is a fear that in some cultures some parents may use this technology to select the sex of their children. A study from Punjab, India (E. E. Booth, M. Verma, and R. S. Beri, "Fetal Sex Determination in Infants in Punjab, India: Correlations and Implications," *BMJ* 309 [12 November 1994]: 1259–1261), reports that, in 1993, in one hospital, 56.9% of the 550 live births that year were boys. It's a medical fact that male babies are slightly more common than female babies. The study's authors report a baseline for this region of 51.7% male live births.

**Question:** Is there evidence that the proportion of male births has changed?



**Plan** State what we want to know.

Define the variables and discuss the  $W$ 's.

**Hypotheses** The null hypothesis makes the claim of no difference from the baseline.

Before seeing the data, we were interested in any change in male births, so the alternative hypothesis is two-sided.

**Model** Think about the assumptions and check the appropriate conditions.

For testing proportions, the conditions are the same ones we had for making confidence intervals, except that we check the **Success/Failure Condition** with the *hypothesized* proportions rather than with the *observed* proportions.

Specify the sampling distribution model.

Tell what test you plan to use.

I want to know whether the proportion of male births has changed from the established baseline of 51.7%. The data are the recorded sexes of the 550 live births from a hospital in Punjab, India, in 1993, collected for a study on fetal sex determination. The parameter of interest,  $p$ , is the proportion of male births:

$$H_0: p = 0.517$$

$$H_A: p \neq 0.517$$

- ✓ **Independence Assumption:** There is no reason to think that the sex of one baby can affect the sex of other babies, so births can reasonably be assumed to be independent with regard to the sex of the child.
- ✓ **Randomization Condition:** The 550 live births are not a random sample, so I must be cautious about any general conclusions. I hope that this is a representative year, and I think that the births at this hospital may be typical of this area of India.
- ✓ **10% Condition:** I would like to be able to make statements about births at similar hospitals in India. These 550 births are fewer than 10% of all of those births.
- ✓ **Success/Failure Condition:** Both  $np_0 = 550(0.517) = 284.35$  and  $nq_0 = 550(0.483) = 265.65$  are greater than 10; I expect the births of at least 10 boys and at least 10 girls, so the sample is large enough.

The conditions are satisfied, so I can use a Normal model and perform a **one-proportion z-test**.

SHOW

**Mechanics** The null model gives us the mean, and (because we are working with proportions) the mean gives us the standard deviation.

We find the z-score for the observed proportion to find out how many standard deviations it is from the hypothesized proportion.

Make a picture. Sketch a Normal model centered at  $p_0 = 0.517$ . Shade the region to the right of the observed proportion, and because this is a two-tail test, also shade the corresponding region in the other tail.

From the z-score, we can find the P-value, which tells us the probability of observing a value that extreme (or more). Use technology or a table (see p. 473.).

Because this is a two-tail test, the P-value is the probability of observing an outcome more than 2.44 standard deviations from the mean of a Normal model *in either direction*. We must therefore *double* the probability we find in the upper tail.

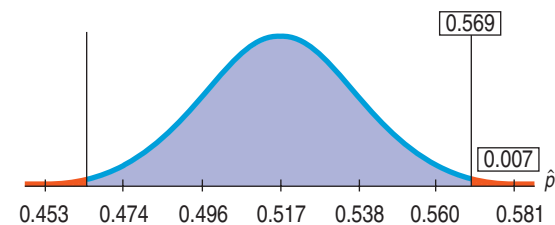
The null model is a Normal distribution with a mean of 0.517 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.517)(1 - 0.517)}{550}} \\ = 0.0213$$

The observed proportion,  $\hat{p}$ , is 0.569, so

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.569 - 0.517}{0.0213} = 2.44$$

The sample proportion lies 2.44 standard deviations above the mean.



$$P = 2P(z > 2.44) = 2(0.0073) = 0.0146$$

TELL

**Conclusion** State your conclusion in context.

This P-value is roughly 1 time in 70. That's clearly significant, but don't jump to other conclusions. We can't be sure how this deviation came about. For instance, we don't know whether this hospital is typical, or whether the time period studied was selected at random.

The P-value of 0.0146 says that if the true proportion of male babies were still at 51.7%, then an observed proportion as different as 56.9% male babies would occur at random only about 15 times in 1000. With a P-value this small, I reject  $H_0$ . This is strong evidence that the birth ratio of boys to girls is not equal to its natural level. It appears that the proportion of boys may have increased.

How big an increase are we talking about? Let's find a confidence interval for the proportion of male births.

THINK

AGAIN

**Model** Check the conditions.

The conditions are identical to those for the hypothesis test, with one difference. Now we are not given a hypothesized proportion,  $p_0$ , so we must instead work with the observed proportion  $\hat{p}$ .

✓ **Success/Failure Condition:** Both  $n\hat{p} = 550(0.569) = 313$  and  $n\hat{q} = 237$  are at least 10.

Specify the sampling distribution model.  
Tell what method you plan to use.

The conditions are satisfied, so I can model the sampling distribution of the proportion with a Normal model and find a **one-proportion z-interval**.



**Mechanics** We can't find the sampling model standard deviation from the null model proportion. (In fact, we've just rejected it.) Instead, we find the standard error of  $\hat{p}$  from the *observed* proportions. Other than that substitution, the calculation looks the same as for the hypothesis test.

With this large a sample size, the difference is negligible, but in smaller samples, it could make a bigger difference.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.569)(1 - 0.569)}{550}} = 0.0211$$

The sampling model is Normal, so for a 95% confidence interval, the critical value  $z^* = 1.96$ .

The margin of error is

$$ME = z^* \times SE(\hat{p}) = 1.96(0.0211) = 0.041$$

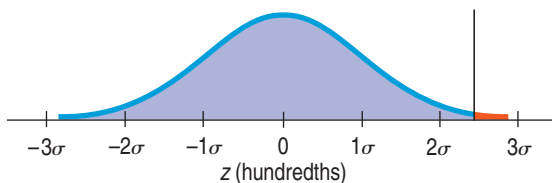
So the 95% confidence interval is

$$0.569 \pm 0.041 \text{ or } (0.528, 0.610).$$



**Conclusion** Confidence intervals help us think about the size of the effect. Here we can see that the change from the baseline of 51.7% male births might be quite substantial.

We are 95% confident that the true proportion of male births is between 52.8% and 61.0%.



Here's a portion of a Normal table that gives the probability we needed for the hypothesis test. At  $z = 2.44$ , the table gives the percentile as 0.9927. The upper-tail probability (shaded red) is, therefore,  $1 - 0.9927 = 0.0073$ ; so, for our two-sided test, the P-value is  $2(0.0073) = 0.0146$ .

z	0.00	0.01	0.02	0.03	0.04	0.05
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960



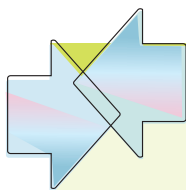
## WHAT CAN GO WRONG?

### Don't We Want to Reject the Null?

Often the folks who collect the data or perform the experiment hope to reject the null. (They hope the new drug is better than the placebo, or new ad campaign is better than the old one.) But when we practice Statistics, we can't allow that hope to affect our decision. The essential attitude for a hypothesis tester is skepticism. Until we become convinced otherwise, we cling to the null's assertion that there's nothing unusual, no effect, no difference, etc. As in a jury trial, the burden of proof rests with the alternative hypothesis—innocent until proven guilty. When you test a hypothesis, you must act as judge and jury, but you are not the prosecutor.

Hypothesis tests are so widely used—and so widely misused—that we've devoted all of the next chapter to discussing the pitfalls involved, but there are a few issues that we can talk about already.

- ▶ **Don't base your null hypotheses on what you see in the data.** You are not allowed to look at the data first and then adjust your null hypothesis so that it will be rejected. When your sample value turns out to be  $\hat{p} = 51.8\%$ , with a standard deviation of 1%, don't form a null hypothesis like  $H_0: p = 49.8\%$ , knowing that you can reject it. You should always *Think* about the situation you are investigating and make your null hypothesis describe the “nothing interesting” or “nothing has changed” scenario. No peeking at the data!
- ▶ **Don't base your alternative hypothesis on the data, either.** Again, you need to *Think* about the situation. Are you interested only in knowing whether something has *increased*? Then write a one-sided (upper-tail) alternative. Or would you be equally interested in a change in either direction? Then you want a two-sided alternative. You should decide whether to do a one- or two-sided test based on what results would be of interest to you, not what you see in the data.
- ▶ **Don't make your null hypothesis what you want to show to be true.** Remember, the null hypothesis is the status quo, the nothing-is-strange-here position a skeptic would take. You wonder whether the data cast doubt on that. You can reject the null hypothesis, but you can never “accept” or “prove” the null.
- ▶ **Don't forget to check the conditions.** The reasoning of inference depends on randomization. No amount of care in calculating a test result can recover from biased sampling. The probabilities we compute depend on the independence assumption. And our sample must be large enough to justify our use of a Normal model.
- ▶ **Don't accept the null hypothesis.** You may not have found enough evidence to reject it, but you surely have *not* proven it's true!
- ▶ **If you fail to reject the null hypothesis, don't think that a bigger sample would be more likely to lead to rejection.** If the results you looked at were “almost” significant, it's enticing to think that because you would have rejected the null had these same observations come from a larger sample, then a larger sample would surely lead to rejection. Don't be misled. Remember, each sample is different, and a larger sample won't necessarily duplicate your current observations. Indeed, the Central Limit Theorem tells us that statistics will vary *less* in larger samples. We should therefore expect such results to be less extreme. Maybe they'd be statistically significant but maybe (perhaps even probably) not. Even if you fail to reject the null hypothesis, it's a good idea to examine a confidence interval. If none of the plausible parameter values in the interval would matter to you (for example, because none would be *practically* significant), then even a larger study with a correspondingly smaller standard error is unlikely to be worthwhile.



## CONNECTIONS

Hypothesis tests and confidence intervals share many of the same concepts. Both rely on sampling distribution models, and because the models are the same and require the same assumptions, both check the same conditions. They also calculate many of the same statistics. Like confidence intervals, hypothesis tests use the standard deviation of the sampling distribution as a ruler, as we first saw in Chapter 6.

For testing, we find ourselves looking once again at z-scores, and we compute the P-value by finding the distance of our test statistic from the center of the null model. P-values are conditional probabilities. They give the probability of observing the result we have seen (or one even more extreme) *given* that the null hypothesis is true.

The Standard Normal model is here again as our connection between z-score values and probabilities.



## WHAT HAVE WE LEARNED?

We've learned to use what we see in a random sample to test a particular hypothesis about the world. This is our second step in statistical inference, complementing our use of confidence intervals.

We've learned that testing a hypothesis involves proposing a model, then seeing whether the data we observe are consistent with that model or are so unusual that we must reject it. We do this by finding a P-value—the probability that data like ours could have occurred if the model is correct.

We've learned that:

- ▶ We start with a null hypothesis specifying the parameter of a model we'll test using our data.
- ▶ Our alternative hypothesis can be one- or two-sided, depending on what we want to learn.
- ▶ We must check the appropriate assumptions and conditions before proceeding with our test.
- ▶ If the data are out of line with the null hypothesis model, the P-value will be small and we will reject the null hypothesis.
- ▶ If the data are consistent with the null hypothesis model, the P-value will be large and we will not reject the null hypothesis.
- ▶ We must always state our conclusion in the context of the original question.

And we've learned that confidence intervals and hypothesis tests go hand in hand in helping us think about models. A hypothesis test makes a yes/no decision about the plausibility of a parameter value. The confidence interval shows us the range of plausible values for the parameter.

### Terms

Null hypothesis	460. The claim being assessed in a hypothesis test is called the null hypothesis. Usually, the null hypothesis is a statement of “no change from the traditional value,” “no effect,” “no difference,” or “no relationship.” For a claim to be a testable null hypothesis, it must specify a value for some population parameter that can form the basis for assuming a sampling distribution for a test statistic.
Alternative hypothesis	460. The alternative hypothesis proposes what we should conclude if we find the null hypothesis to be unlikely.
Two-sided alternative (Two-tailed alternative)	466. An alternative hypothesis is two-sided ( $H_A: p \neq p_0$ ) when we are interested in deviations in <i>either</i> direction away from the hypothesized parameter value.
One-sided alternative (One-tailed alternative)	466. An alternative hypothesis is one-sided (e.g., $H_A: p > p_0$ or $H_A: p < p_0$ ) when we are interested in deviations in <i>only one</i> direction away from the hypothesized parameter value.
P-value	461. The probability of observing a value for a test statistic at least as far from the hypothesized value as the statistic value actually observed if the null hypothesis is true. A small P-value indicates either that the observation is improbable or that the probability calculation was based on incorrect assumptions. The assumed truth of the null hypothesis is the assumption under suspicion.
One-proportion z-test	464. A test of the null hypothesis that the proportion of a single sample equals a specified value ( $H_0: p = p_0$ ) by referring the statistic $z = \frac{\hat{p} - p_0}{SD(\hat{p})}$ to a Standard Normal model.

### Skills

THINK

- ▶ Be able to state the null and alternative hypotheses for a one-proportion z-test.
- ▶ Know the conditions that must be true for a one-proportion z-test to be appropriate, and know how to examine your data for violations of those conditions.
- ▶ Be able to identify and use the alternative hypothesis when testing hypotheses. Understand how to choose between a one-sided and two-sided alternative hypothesis, and be able to explain your choice.

SHOW

- ▶ Be able to perform a one-proportion z-test.

TELL

- ▶ Be able to write a sentence interpreting the results of a one-proportion z-test.
- ▶ Know how to interpret the meaning of a P-value in nontechnical language, making clear that the probability claim is made about computed values under the assumption that the null model is true and not about the population parameter of interest.

## HYPOTHESIS TESTS FOR PROPORTIONS ON THE COMPUTER

Hypothesis tests for proportions are so easy and natural that many statistics packages don't offer special commands for them. Most statistics programs want to know the "success" and "failure" status for each case. Usually these are given as 1 or 0, but they might be category names like "yes" and "no." Often we just know the proportion of successes,  $\hat{p}$ , and the total count,  $n$ . Computer packages don't usually deal naturally with summary data like this, but the statistics routines found on many graphing calculators do. These calculators allow you to test hypotheses from summaries of the data—usually, all you need to enter are the number of successes and the sample size.

## EXERCISES

- Hypotheses.** Write the null and alternative hypotheses you would use to test each of the following situations:
  - A governor is concerned about his "negatives"—the percentage of state residents who express disapproval of his job performance. His political committee pays for a series of TV ads, hoping that they can keep the negatives below 30%. They will use follow-up polling to assess the ads' effectiveness.
  - Is a coin fair?
  - Only about 20% of people who try to quit smoking succeed. Sellers of a motivational tape claim that listening to the recorded messages can help people quit.
- More hypotheses.** Write the null and alternative hypotheses you would use to test each situation.
  - In the 1950s only about 40% of high school graduates went on to college. Has the percentage changed?
  - 20% of cars of a certain model have needed costly transmission work after being driven between 50,000 and 100,000 miles. The manufacturer hopes that a redesign of a transmission component has solved this problem.
  - We field-test a new-flavor soft drink, planning to market it only if we are sure that over 60% of the people like the flavor.
- Negatives.** After the political ad campaign described in Exercise 1a, pollsters check the governor's negatives. They test the hypothesis that the ads produced no change against the alternative that the negatives are now below 30% and find a P-value of 0.22. Which conclusion is appropriate? Explain.
  - There's a 22% chance that the ads worked.
  - There's a 78% chance that the ads worked.
  - There's a 22% chance that their poll is correct.
  - There's a 22% chance that natural sampling variation could produce poll results like these if there's really no change in public opinion.
- Dice.** The seller of a loaded die claims that it will favor the outcome 6. We don't believe that claim, and roll the die 200 times to test an appropriate hypothesis. Our P-value turns out to be 0.03. Which conclusion is appropriate? Explain.
  - There's a 3% chance that the die is fair.
  - There's a 97% chance that the die is fair.
  - There's a 3% chance that a loaded die could randomly produce the results we observed, so it's reasonable to conclude that the die is fair.
  - There's a 3% chance that a fair die could randomly produce the results we observed, so it's reasonable to conclude that the die is loaded.
- Relief.** A company's old antacid formula provided relief for 70% of the people who used it. The company tests a new formula to see if it is better and gets a P-value of 0.27. Is it reasonable to conclude that the new formula and the old one are equally effective? Explain.
- Cars.** A survey investigating whether the proportion of today's high school seniors who own their own cars is higher than it was a decade ago finds a P-value of 0.017. Is it reasonable to conclude that more high-schoolers have cars? Explain.
- He cheats!** A friend of yours claims that when he tosses a coin he can control the outcome. You are skeptical and want him to prove it. He tosses the coin, and you call heads; it's tails. You try again and lose again.
  - Do two losses in a row convince you that he really can control the toss? Explain.
  - You try a third time, and again you lose. What's the probability of losing three tosses in a row if the process is fair?
  - Would three losses in a row convince you that your friend cheats? Explain.
  - How many times in a row would you have to lose in order to be pretty sure that this friend really can control the toss? Justify your answer by calculating a probability and explaining what it means.
- Candy.** Someone hands you a box of a dozen chocolate-covered candies, telling you that half are vanilla creams and the other half peanut butter. You pick candies at random and discover the first three you eat are all vanilla.

- a) If there really were 6 vanilla and 6 peanut butter candies in the box, what is the probability that you would have picked three vanillas in a row?
- b) Do you think there really might have been 6 of each? Explain.
- c) Would you continue to believe that half are vanilla if the fourth one you try is also vanilla? Explain.

9. **Cell phones.** Many people have trouble setting up all the features of their cell phones, so a company has developed what it hopes will be easier instructions. The goal is to have at least 96% of customers succeed. The company tests the new system on 200 people, of whom 188 were successful. Is this strong evidence that the new system fails to meet the company's goal? A student's test of this hypothesis is shown. How many mistakes can you find?

$$H_0: \hat{p} = 0.96$$

$$H_A: \hat{p} \neq 0.96$$

$$\text{SRS}, 0.96(200) > 10$$

$$\frac{188}{200} = 0.94; \quad SD(\hat{p}) = \sqrt{\frac{(0.94)(0.06)}{200}} = 0.017$$

$$z = \frac{0.96 - 0.94}{0.017} = 1.18$$

$$P = P(z > 1.18) = 0.12$$

There is strong evidence the new instructions don't work.

10. **Got milk?** In November 2001, the *Ag Globe Trotter* newsletter reported that 90% of adults drink milk. A regional farmers' organization planning a new marketing campaign across its multicounty area polls a random sample of 750 adults living there. In this sample, 657 people said that they drink milk. Do these responses provide strong evidence that the 90% figure is not accurate for this region? Correct the mistakes you find in a student's attempt to test an appropriate hypothesis.

$$H_0: \hat{p} = 0.9$$

$$H_A: \hat{p} < 0.9$$

$$\text{SRS}, 750 > 10$$

$$\frac{657}{750} = 0.876; \quad SD(\hat{p}) = \sqrt{\frac{(0.88)(0.12)}{750}} = 0.012$$

$$z = \frac{0.876 - 0.90}{0.012} = -2$$

$$P = P(z > -2) = 0.977$$

There is more than a 97% chance that the stated percentage is correct for this region.

11. **Dowsing.** In a rural area, only about 30% of the wells that are drilled find adequate water at a depth of 100 feet or less. A local man claims to be able to find water by "dowsing"—using a forked stick to indicate where the well should be drilled. You check with 80 of his customers and find that 27 have wells less than 100 feet deep. What do you conclude about his claim?
- Write appropriate hypotheses.
  - Check the necessary assumptions.
  - Perform the mechanics of the test. What is the P-value?
  - Explain carefully what the P-value means in context.
  - What's your conclusion?

12. **Abnormalities.** In the 1980s it was generally believed that congenital abnormalities affected about 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an increase in the incidence of abnormalities. A recent study examined 384 children and found that 46 of them showed signs of an abnormality. Is this strong evidence that the risk has increased?

- Write appropriate hypotheses.
- Check the necessary assumptions.
- Perform the mechanics of the test. What is the P-value?
- Explain carefully what the P-value means in context.
- What's your conclusion?
- Do environmental chemicals cause congenital abnormalities?

13. **Absentees.** The National Center for Education Statistics monitors many aspects of elementary and secondary education nationwide. Their 1996 numbers are often used as a baseline to assess changes. In 1996 34% of students had not been absent from school even once during the previous month. In the 2000 survey, responses from 8302 students showed that this figure had slipped to 33%. Officials would, of course, be concerned if student attendance were declining. Do these figures give evidence of a change in student attendance?

- Write appropriate hypotheses.
- Check the assumptions and conditions.
- Perform the test and find the P-value.
- State your conclusion.
- Do you think this difference is meaningful? Explain.

14. **Educated mothers.** The National Center for Education Statistics monitors many aspects of elementary and secondary education nationwide. Their 1996 numbers are often used as a baseline to assess changes. In 1996, 31% of students reported that their mothers had graduated from college. In 2000, responses from 8368 students found that this figure had grown to 32%. Is this evidence of a change in education level among mothers?

- Write appropriate hypotheses.
- Check the assumptions and conditions.
- Perform the test and find the P-value.
- State your conclusion.
- Do you think this difference is meaningful? Explain.

15. **Contributions, please, part II.** In Exercise 19.15 you learned that the Paralyzed Veterans of America is a philanthropic organization that relies on contributions. They send free mailing labels and greeting cards to potential donors on their list and ask for a voluntary contribution. To test a new campaign, the organization recently sent letters to a random sample of 100,000 potential donors and received 4781 donations. They've had a contribution rate of 5% in past campaigns, but a staff member worries that the rate will be lower if they run this campaign as currently designed.

- What are the hypotheses?
- Are the assumptions and conditions for inference met?
- Do you think the rate would drop? Explain.

16. **Take the offer, part II.** In Exercise 19.16 you learned that First USA, a major credit card company, is planning a new offer for their current cardholders. First USA will give double airline miles on purchases for the next 6 months if the cardholder goes online and registers for this offer. To test the effectiveness of this campaign, the company recently sent out offers to a random sample of 50,000 cardholders. Of those, 1184 registered. A staff member suspects that the success rate for the full campaign will be comparable to the standard 2% rate that they are used to seeing in similar campaigns. What do you predict?
- What are the hypotheses?
  - Are the assumptions and conditions for inference met?
  - Do you think the rate would change if they use this fundraising campaign? Explain.
17. **Law School.** According to the Law School Admission Council, in the fall of 2006, 63% of law school applicants were accepted to some law school.<sup>4</sup> The training program *LSATisfaction* claims that 163 of the 240 students trained in 2006 were admitted to law school. You can safely consider these trainees to be representative of the population of law school applicants. Has *LSATisfaction* demonstrated a real improvement over the national average?
- What are the hypotheses?
  - Check the conditions and find the P-value.
  - Would you recommend this program based on what you see here? Explain.
18. **Med School.** According to the Association of American Medical Colleges, only 46% of medical school applicants were admitted to a medical school in the fall of 2006.<sup>5</sup> Upon hearing this, the trustees of Striving College expressed concern that only 77 of the 180 students in their class of 2006 who applied to medical school were admitted. The college president assured the trustees that this was just the kind of year-to-year fluctuation in fortunes that is to be expected and that, in fact, the school's success rate was consistent with the national average. Who is right?
- What are the hypotheses?
  - Check the conditions and find the P-value.
  - Are the trustees right to be concerned, or is the president correct? Explain.
19. **Pollution.** A company with a fleet of 150 cars found that the emissions systems of 7 out of the 22 they tested failed to meet pollution control guidelines. Is this strong evidence that more than 20% of the fleet might be out of compliance? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
20. **Scratch and dent.** An appliance manufacturer stockpiles washers and dryers in a large warehouse for shipment to retail stores. Sometimes in handling them the appliances get damaged. Even though the damage may be minor, the company must sell those machines at drastically reduced prices. The company goal is to keep the level of damaged machines below 2%. One day an inspector randomly checks 60 washers and finds that 5 of them have scratches or dents. Is this strong evidence that the warehouse is failing to meet the company goal? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
21. **Twins.** In 2001 a national vital statistics report indicated that about 3% of all births produced twins. Is the rate of twin births the same among very young mothers? Data from a large city hospital found that only 7 sets of twins were born to 469 teenage girls. Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
22. **Football 2006.** During the 2006 season, the home team won 136 of the 240 regular-season National Football League games. Is this strong evidence of a home field advantage in professional football? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
23. **WebZine.** A magazine is considering the launch of an online edition. The magazine plans to go ahead only if it's convinced that more than 25% of current readers would subscribe. The magazine contacted a simple random sample of 500 current subscribers, and 137 of those surveyed expressed interest. What should the company do? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
24. **Seeds.** A garden center wants to store leftover packets of vegetable seeds for sale the following spring, but the center is concerned that the seeds may not germinate at the same rate a year later. The manager finds a packet of last year's green bean seeds and plants them as a test. Although the packet claims a germination rate of 92%, only 171 of 200 test seeds sprout. Is this evidence that the seeds have lost viability during a year in storage? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
25. **Women executives.** A company is criticized because only 13 of 43 people in executive-level positions are women. The company explains that although this proportion is lower than it might wish, it's not surprising given that only 40% of all its employees are women. What do you think? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed.
26. **Jury.** Census data for a certain county show that 19% of the adult residents are Hispanic. Suppose 72 people are called for jury duty and only 9 of them are Hispanic. Does this apparent underrepresentation of Hispanics call into question the fairness of the jury selection system? Explain.

<sup>4</sup> As reported by the Cornell office of career services in their *Class of 2006 Postgraduate Report*.

<sup>5</sup> *Ibid.*

27. **Dropouts.** Some people are concerned that new tougher standards and high-stakes tests adopted in many states have driven up the high school dropout rate. The National Center for Education Statistics reported that the high school dropout rate for the year 2004 was 10.3%. One school district whose dropout rate has always been very close to the national average reports that 210 of their 1782 high school students dropped out last year. Is this evidence that their dropout rate may be increasing? Explain.
28. **Acid rain.** A study of the effects of acid rain on trees in the Hopkins Forest shows that 25 of 100 trees sampled exhibited some sort of damage from acid rain. This rate seemed to be higher than the 15% quoted in a recent *Environmetrics* article on the average proportion of damaged trees in the Northeast. Does the sample suggest that trees in the Hopkins Forest are more susceptible than trees from the rest of the region? Comment, and write up your own conclusions based on an appropriate confidence interval as well as a hypothesis test. Include any assumptions you made about the data.
29. **Lost luggage.** An airline's public relations department says that the airline rarely loses passengers' luggage. It further claims that on those occasions when luggage is lost, 90% is recovered and delivered to its owner within 24 hours. A consumer group that surveyed a large number of air travelers found that only 103 of 122 people who lost luggage on that airline were reunited with the missing items by the next day. Does this cast doubt on the airline's claim? Explain.
30. **TV ads.** A start-up company is about to market a new computer printer. It decides to gamble by running commercials during the Super Bowl. The company hopes that name recognition will be worth the high cost of the ads. The goal of the company is that over 40% of the public recognize its brand name and associate it with computer equipment. The day after the game, a pollster contacts 420 randomly chosen adults and finds that 181 of them know that this company manufactures printers. Would you recommend that the company continue to advertise during Super Bowls? Explain.
31. **John Wayne.** Like a lot of other Americans, John Wayne died of cancer. But is there more to this story? In 1955 Wayne was in Utah shooting the film *The Conqueror*. Across the state line, in Nevada, the United States military was testing atomic bombs. Radioactive fallout from those tests drifted across the filming location. A total of 46 of the 220 people working on the film eventually died of cancer. Cancer experts estimate that one would expect only about 30 cancer deaths in a group this size.
- a) Is the death rate among the movie crew unusually high?
- b) Does this prove that exposure to radiation increases the risk of cancer?
32. **AP Stats.** The College Board reported that 60% of all students who took the 2006 AP Statistics exam earned scores of 3 or higher. One teacher wondered if the performance of her school was different. She believed that year's students to be typical of those who will take AP Stats at that school and was pleased when 65% of her 54 students achieved scores of 3 or better. Can she claim that her school is different? Explain.



### JUST CHECKING Answers

1. You can't conclude that the null hypothesis is true. You can conclude only that the experiment was unable to reject the null hypothesis. They were unable, on the basis of 12 patients, to show that aspirin was effective.
2. The null hypothesis is  $H_0: p = 0.75$ .
3. With a P-value of 0.0001, this is very strong evidence against the null hypothesis. We can reject  $H_0$  and conclude that the improved version of the drug gives relief to a higher proportion of patients.
4. The parameter of interest is the proportion,  $p$ , of all delinquent customers who will pay their bills.  $H_0: p = 0.30$  and  $H_A: p > 0.30$ .
5. The very low P-value leads us to reject the null hypothesis. There is strong evidence that the DVD is more effective in getting people to start paying their debts than just sending a letter had been.
6. All we know is that there is strong evidence to suggest that  $p > 0.30$ . We don't know how much higher than 30% the new proportion is. We'd like to see a confidence interval to see if the new method is worth the cost.

## More About Tests and Intervals



**WHO** Florida motorcycle riders aged 20 and younger involved in motorcycle accidents

**WHAT** % wearing helmets

**WHEN** 2001–2003

**WHERE** Florida

**WHY** Assessment of injury rates commissioned by the National Highway Traffic Safety Administration (NHTSA)

In 2000 Florida changed its motorcycle helmet law. No longer are riders 21 and older required to wear helmets. Under the new law, those under 21 still must wear helmets, but a report by the Preusser Group ([www.preussergroup.com](http://www.preussergroup.com)) suggests that helmet use may have declined in this group, too.

It isn't practical to survey young motorcycle riders. (For example, how can you construct a sampling frame? If you contacted licensed riders, would they admit to riding illegally without a helmet?) The researchers adopted a different strategy. Police reports of motorcycle accidents record whether the rider wore a helmet and give the rider's age. Before the change in the helmet law, 60% of youths involved in a motorcycle accident had been wearing their helmets. The Preusser study looked at accident reports during 2001–2003, the three years following the law change, considering these riders to be a representative sample of the larger population. They observed 781 young riders who were involved in accidents. Of these, 396 (or 50.7%) were wearing helmets. Is this evidence of a decline in helmet-wearing, or just the natural fluctuation of such statistics?

## Zero In on the Null

Null hypotheses have special requirements. In order to perform a statistical test of the hypothesis, the null must be a statement about the value of a parameter for a model. We use this value to compute the probability that the observed sample statistic—or something even farther from the null value—might occur.

How do we choose the null hypothesis? The appropriate null arises directly from the context of the problem. It is dictated, not by the data, but by the situation. One good way to identify both the null and alternative hypotheses is to think about the *Why* of the situation. Typical null hypotheses might be that the proportion of patients recovering after receiving a new drug is the same as we would expect of patients receiving a placebo or that the mean strength attained by athletes training with new equipment is the same as with the old equipment. The alternative hypotheses would be that the new drug cures a higher proportion of patients or that the new equipment results in a greater mean strength.

To write a null hypothesis, you can't just choose any parameter value you like. The null must relate to the question at hand. Even though the null usually means no difference or no change, you can't automatically interpret "null" to mean zero. A claim that "nobody" wears a motorcycle helmet would be absurd. The null hypothesis for the Florida study could be that the true rate of helmet use remained the same among young riders after the law changed. You need to find the value for the parameter in the null hypothesis from the context of the problem.

There is a temptation to state your *claim* as the null hypothesis. As we have seen, however, you cannot prove a null hypothesis true any more than you can prove a defendant innocent. So, it makes more sense to use what you want to show as the *alternative*. This way, if you reject the null, you are left with what you want to show.

## FOR EXAMPLE

### Writing hypotheses

The diabetes drug Avandia<sup>®</sup> was approved to treat Type 2 diabetes in 1999. But in 2007 an article in the *New England Journal of Medicine (NEJM)*<sup>1</sup> raised concerns that the drug might carry an increased risk of heart attack. This study combined results from a number of other separate studies to obtain an overall sample of 4485 diabetes patients taking Avandia. People with Type 2 diabetes are known to have about a 20.2% chance of suffering a heart attack within a seven-year period. According to the article's author, Dr. Steven E. Nissen,<sup>2</sup> the risk found in the *NEJM* study was equivalent to a 28.9% chance of heart attack over seven years. The FDA is the government agency responsible for relabeling Avandia to warn of the risk if it is judged to be unsafe. Although the statistical methods they use are more sophisticated, we can get an idea of their reasoning with the tools we have learned.

**Question:** What null hypothesis and alternative hypothesis about seven-year heart attack risk would you test? Explain.

$$H_0: p = 0.202$$

$$H_A: p > 0.202$$

The parameter of interest is the proportion of diabetes patients suffering a heart attack in seven years. The FDA is concerned only with whether Avandia increases the seven-year risk of heart attacks above the baseline value of 20.2%, so a one-sided upper-tail test is appropriate.

**One-sided or two?** In the 1930s, a series of experiments was performed at Duke University in an attempt to see whether humans were capable of extrasensory perception, or ESP. Psychologist Karl Zener designed a set of cards with 5 symbols, later made infamous in the movie *Ghostbusters*:



In the experiment, the "sender" selects one of the 5 cards at random from a deck and then concentrates on it. The "receiver" tries to determine which card it is. If we let  $p$  be the proportion of correct responses, what's the null hypothesis? The null hypothesis is that ESP makes no difference. Without ESP, the receiver would just be guessing, and since there are 5 possible responses, there would be a 20% chance of guessing each card correctly. So,  $H_0$  is  $p = 0.20$ . What's the alternative? It seems that it should be  $p > 0.20$ , a one-sided alternative. But some ESP researchers have expressed the claim that if the proportion guessed were much *lower* than expected, that would show an "interference" and should be considered evidence for ESP as well. So they argue for a two-sided alternative.

<sup>1</sup> Steven E. Nissen, M.D., and Kathy Wolski, M.P.H., "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes," *NEJM* 2007; 356.

<sup>2</sup> Interview reported in the *New York Times* [May 26, 2007].



## STEP-BY-STEP EXAMPLE

## Another One-Proportion z-Test

Let's try to answer the question raised at the start of the chapter.

**Question:** Has helmet use in Florida declined among riders under the age of 21 subsequent to the change in the helmet laws?

THINK

**Plan** State the problem and discuss the variables and the  $W$ 's.

**Hypotheses** The null hypothesis is established by the rate set before the change in the law. The study was concerned with safety, so they'll want to know of any decline in helmet use, making this a lower-tail test.

I want to know whether the rate of helmet wearing among Florida's motorcycle riders under the age of 21 remained at 60% after the law changed to allow older riders to go without helmets. I have data from accident records showing 396 of 781 young riders were wearing helmets.

$$H_0: p = 0.60$$

$$H_A: p < 0.60$$

SHOW

**Model** Check the conditions.

The Risky Behavior Surveillance survey is in fact a complex, multistage sample, but it is randomized and great effort is taken to make it representative. It is safe to treat it as though it were a random sample.

Specify the sampling distribution model and name the test.

- ✓ **Independence Assumption:** The data are for riders involved in accidents during a three-year period. Individuals are independent of one another.
- ✗ **Randomization Condition:** No randomization was applied, but we are considering these riders involved in accidents to be a representative sample of all riders. We should take care in generalizing our conclusions.
- ✓ **10% Condition:** These 781 riders are a small sample of a larger population of all young motorcycle riders.
- ✓ **Success/Failure Condition:** We'd expect  $np = 781(0.6) = 468.6$  helmeted riders and  $nq = 781(0.4) = 312.4$  non-helmeted. Both are at least 10.

The conditions are satisfied, so I can use a Normal model and perform a **one-proportion z-test**.

SHOW

**Mechanics** Find the standard deviation of the sampling model using the hypothesized proportion.

Find the z-score for the observed proportion.

There were 396 helmet wearers among the 781 accident victims.

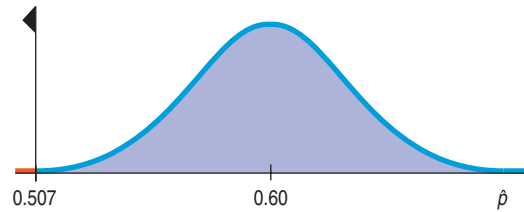
$$\hat{p} = \frac{396}{781} = 0.507$$

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.60)(0.40)}{781}} = 0.0175$$

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.507 - 0.60}{0.0175} = -5.31$$

Make a picture. Sketch a Normal model centered at the hypothesized helmet rate of 60%. This is a lower-tail test, so shade the region to the left of the observed rate.

Given this z-score, the P-value is obviously very low.



The observed helmet rate is 5.31 standard deviations below the former rate. The corresponding P-value is less than 0.001.



**Conclusion** Link the P-value to your decision about the null hypothesis, and then state your conclusion in context.

The very small P-value says that if the true rate of helmet-wearing among riders under 21 were still 60%, the probability of observing a rate no higher than 50.7% in a sample like this is less than 1 chance in 1000, so I reject the null hypothesis. There is strong evidence that there has been a decline in helmet use among riders under 21.

## How to Think About P-values

### Which Conditional?

Suppose that as a political science major you are offered the chance to be a White House intern. There would be a very high probability that next summer you'd be in Washington, D.C. That is,  $P(\text{Washington} | \text{Intern})$  would be high. But if we find a student in Washington, D.C., is it likely that he's a White House intern? Almost surely not;  $P(\text{Intern} | \text{Washington})$  is low. You can't switch around conditional probabilities. The P-value is  $P(\text{data} | H_0)$ . We might wish we could report  $P(H_0 | \text{data})$ , but these two quantities are NOT the same.

A P-value actually is a conditional probability. It tells us the probability of getting results at least as unusual as the observed statistic, *given* that the null hypothesis is true. We can write  $P\text{-value} = P(\text{observed statistic value [or even more extreme]} | H_0)$ .

Writing the P-value this way helps to make clear that the P-value is *not* the probability that the null hypothesis is true. It is a probability about the data. Let's say that again:

*The P-value is not the probability that the null hypothesis is true.*

The P-value is not even the conditional probability that the null hypothesis is true given the data. We would write that probability as  $P(H_0 | \text{observed statistic value})$ . This is a conditional probability but in reverse. It would be nice to know this, but it's impossible to calculate without making additional assumptions. As we saw in Chapter 15, reversing the order in a conditional probability is difficult, and the results can be counterintuitive.

We can find the P-value,  $P(\text{observed statistic value} | H_0)$ , because  $H_0$  gives the parameter values that we need to find the required probability. But there's no direct way to find  $P(H_0 | \text{observed statistic value})$ .<sup>3</sup> As tempting as it may be to say that a P-value of 0.03 means there's a 3% chance that the null hypothesis is true, that just isn't right. All we can say is that, given the null hypothesis, there's a 3% chance of observing the statistic value that we have actually observed (or one more unlike the null value).

<sup>3</sup> The approach to statistical inference known as Bayesian Statistics addresses the question in just this way, but it requires more advanced mathematics and more assumptions. See p. 358 for more about the founding father of this approach.

*“The wise man proportions his belief to the evidence.”*

—David Hume,  
“Enquiry Concerning Human Understanding,” 1748

*“You’re so guilty now.”*

—Rearview Mirror

**How guilty is the suspect?** We might like to know  $P(H_0 | \text{data})$ , but when you think about it, we can’t talk about the probability that the null hypothesis is true. The null is not a random event, so either it is true or it isn’t. The data, however, are random in the sense that if we were to repeat a randomized experiment or draw another random sample, we’d get different data and expect to find a different statistic value. So we can talk about the probability of the data given the null hypothesis, and that’s the P-value.

But it does make sense that the smaller the P-value, the more confident we can be in declaring that we doubt the null hypothesis. Think again about the jury trial. Our null hypothesis is that the defendant is innocent. Then the evidence starts rolling in. A car the same color as his was parked in front of the bank. Well, there are lots of cars that color. The probability of that happening (given his innocence) is pretty high, so we’re not persuaded that he’s guilty. The bank’s security camera showed the robber was male and about the defendant’s height and weight. Hmm. Could that be a coincidence? If he’s innocent, then it’s a little less likely that the car and description would *both* match, so our P-value goes down. We’re starting to question his innocence a little. Witnesses said the robber wore a blue jacket just like the one the police found in a garbage can behind the defendant’s house. Well, if he’s innocent, then that doesn’t seem very likely, does it? If he’s really innocent, the probability that all of these could have happened is getting pretty low. Now our P-value may be small enough to be called “beyond a reasonable doubt” and lead to a conviction. Each new piece of evidence strains our skepticism a bit more. The more compelling the evidence—the more *unlikely* it would be were he innocent—the more convinced we become that he’s guilty.

But even though it may make *us* more confident in declaring him guilty, additional evidence does not make *him* any guiltier. Either he robbed the bank or he didn’t. Additional evidence (like the teller picking him out of a police lineup) just makes us more confident that we did the right thing when we convicted him. The lower the P-value, the more comfortable we feel about our decision to reject the null hypothesis, but the null hypothesis doesn’t get any more false.

## FOR EXAMPLE

### Thinking about the P-value

**Recap:** A *New England Journal of Medicine* paper reported that the seven-year risk of heart attack in diabetes patients taking the drug Avandia was increased from the baseline of 20.2% to an estimated risk of 28.9% and said the P-value was 0.03.

**Question:** How should the P-value be interpreted in this context?

The P-value =  $P(\hat{p} \geq 28.9\% | p = 20.2\%)$ . That is, it’s the probability of seeing such a high heart attack rate among the people studied if, in fact, taking Avandia really didn’t increase the risk at all.

## What to Do with a High P-value



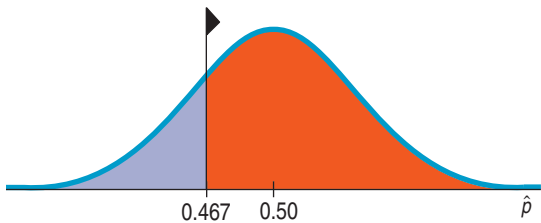
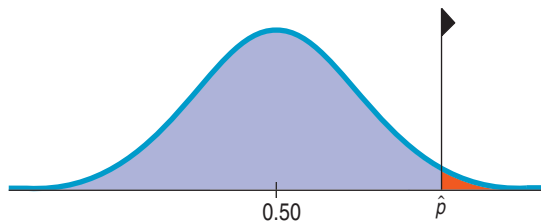
Therapeutic touch (TT), taught in many schools of nursing, is a therapy in which the practitioner moves her hands near, but does not touch, a patient in an attempt to manipulate a “human energy field.” Therapeutic touch practitioners believe that by adjusting this field they can promote healing. However, no instrument has ever detected a human energy field, and no experiment has ever shown that TT practitioners can detect such a field.

In 1998, the *Journal of the American Medical Association* published a paper reporting work by a then nine-year-old girl.<sup>4</sup> She had performed a simple experiment in

<sup>4</sup> L. Rosa, E. Rosa, L. Sarner, and S. Barrett, “A Close Look at Therapeutic Touch,” *JAMA* 279(13) [1 April 1998]: 1005–1010.

**A S** **Video: Is There Evidence for Therapeutic Touch?** This video shows the experiment and tells the story.

**A S** **Activity: Testing Therapeutic Touch.** Perform the one-proportion z-test using *ActivStats* technology. The test in *ActivStats* is two-sided. Do you think this is the appropriate choice?



which she challenged 15 TT practitioners to detect whether her unseen hand was hovering over their left or right hand (selected by the flip of a coin).

The practitioners “warmed up” with a period during which they could see the experimenter’s hand, and each said that they could detect the girl’s human energy field. Then a screen was placed so that the practitioners could not see the girl’s hand, and they attempted 10 trials each. Overall, of 150 trials, the TT practitioners were successful 70 times, for a success proportion of 46.7%. Is there evidence from this experiment that TT practitioners can successfully detect a “human energy field”?

When we see a small P-value, we could continue to believe the null hypothesis and conclude that we just witnessed a rare event. But instead, we trust the data and use it as evidence to reject the null hypothesis.

In the therapeutic touch example, the null hypothesis is that the practitioners are guessing, so we expect them to be right about half the time by chance. That’s why we say  $H_0: p = 0.5$ . They claim that they can detect a “human energy field” and that their success rate should be well above chance, so our alternative is that they would do *better* than guessing. That’s a one-sided alternative hypothesis:  $H_A: p > 0.5$ . With a one-sided hypothesis, our P-value is the probability the practitioners could achieve the observed number of successes or *more* even if they were just guessing.

If the practitioners had been highly successful, that would have been unusually lucky for guessing, so we would have seen a correspondingly low P-value. Since we don’t believe in rare events, we would then have concluded that they weren’t guessing.

But that’s not what happened. What we actually observed was that they did slightly *worse* than 50%, with a  $\hat{p} = 0.467$  success rate.

As the figure shows, the probability of a success rate of 0.467 or *more* is even bigger than 0.5. In this case, it turns out to be 0.793. Obviously, we won’t be rejecting the null hypothesis; for us to reject it, the P-value would have to be quite small. But a P-value of 0.788 seems so big it is almost awkward. With a success rate even lower than chance, we could have concluded right away that we have no evidence for rejecting  $H_0$ .

Big P-values just mean that what we’ve observed isn’t surprising. That is, the results are in line with our assumption that the null hypothesis models the world, so we have no reason to reject it. A big P-value doesn’t prove that the null hypothesis is true, but it certainly offers no evidence that it’s *not* true. When we see a large P-value, all we can say is that we “don’t reject the null hypothesis.”

## FOR EXAMPLE

### More about P-values

**Recap:** The question of whether the diabetes drug Avandia increased the risk of heart attack was raised by a study in the *New England Journal of Medicine*. This study estimated the seven-year risk of heart attack to be 28.9% and reported a P-value of 0.03 for a test of whether this risk was higher than the baseline seven-year risk of 20.2%. An earlier study (the ADOPT study) had estimated the seven-year risk to be 26.9% and reported a P-value of 0.27.

**Question:** Why did the researchers in the ADOPT study not express alarm about the increased risk they had seen?

A P-value of 0.27 means that a heart attack rate at least as high as the one they observed could be expected in 27% of similar experiments even if, in fact, there were no increased risk from taking Avandia. That’s not remarkable enough to reject the null hypothesis. In other words, the ADOPT study wasn’t convincing.

## Alpha Levels

**A S** **Activity: Rejecting the Null Hypothesis.** See alpha levels at work in the animated hypothesis-testing tool.

### NOTATION ALERT:

The first Greek letter,  $\alpha$ , is used in Statistics for the threshold value of a hypothesis test. You'll hear it referred to as the alpha level. Common values are 0.10, 0.05, 0.01, and 0.001.



Sir Ronald Fisher (1890–1962) was one of the founders of modern Statistics.

### It Could Happen to You!

Of course, if the null hypothesis is true, no matter what alpha level you choose, you still have a probability  $\alpha$  of rejecting the null hypothesis by mistake. This is the rare event we want to protect ourselves against. When we do reject the null hypothesis, no one ever thinks that *this* is one of those rare times. As statistician Stu Hunter notes, “The statistician says ‘rare events do happen—but not to me!’”

Sometimes we need to make a firm decision about whether or not to reject the null hypothesis. A jury must *decide* whether the evidence reaches the level of “beyond a reasonable doubt.” A business must *select* a Web design. You need to decide which section of Statistics to enroll in.

When the P-value is small, it tells us that our data are rare, *given the null hypothesis*. As humans, we are suspicious of rare events. If the data are “rare enough,” we just don’t think that could have happened due to chance. Since the data *did* happen, something must be wrong. All we can do now is reject the null hypothesis.

But how rare is “rare”?

We can define “rare event” arbitrarily by setting a threshold for our P-value. If our P-value falls below that point, we’ll reject the null hypothesis. We call such results **statistically significant**. The threshold is called an **alpha level**. Not surprisingly, it’s labeled with the Greek letter  $\alpha$ . Common  $\alpha$  levels are 0.10, 0.05, and 0.01. You have the option—almost the *obligation*—to consider your alpha level carefully and choose an appropriate one for the situation. If you’re assessing the safety of air bags, you’ll want a low alpha level; even 0.01 might not be low enough. If you’re just wondering whether folks prefer their pizza with or without pepperoni, you might be happy with  $\alpha = 0.10$ . It can be hard to justify your choice of  $\alpha$ , though, so often we arbitrarily choose 0.05. Note, however: You must select the alpha level *before* you look at the data. Otherwise you can be accused of cheating by tuning your alpha level to suit the data.

**Where did the value 0.05 come from?** In 1931, in a famous book called *The Design of Experiments*, Sir Ronald Fisher discussed the amount of evidence needed to reject a null hypothesis. He said that it was *situation dependent*, but remarked, somewhat casually, that for many scientific applications, 1 out of 20 *might* be a reasonable value. Since then, some people—indeed some entire disciplines—have treated the number 0.05 as sacrosanct.

The alpha level is also called the **significance level**. When we reject the null hypothesis, we say that the test is “significant at that level.” For example, we might say that we reject the null hypothesis “at the 5% level of significance.”

What can you say if the P-value does not fall below  $\alpha$ ?

When you have not found sufficient evidence to reject the null according to the standard you have established, you should say that “The data have failed to provide sufficient evidence to reject the null hypothesis.” Don’t say that you “accept the null hypothesis.” You certainly haven’t proven or established it; it was merely assumed to begin with. Say that you’ve failed to reject it.

Think again about the therapeutic touch example. The P-value was 0.788. This is so much larger than any reasonable alpha level that we can’t reject  $H_0$ . For this test, we’d conclude, “We fail to reject the null hypothesis. There is insufficient evidence to conclude that the practitioners are performing better than they would if they were just guessing.”

The automatic nature of the reject/fail-to-reject decision when we use an alpha level may make you uncomfortable. If your P-value falls just slightly above your alpha level, you’re not allowed to reject the null. Yet a P-value just barely below the alpha level leads to rejection. If this bothers you, you’re in good company. Many statisticians think it better to report the P-value than to base a decision on an arbitrary alpha level.

**It's in the stars** Some disciplines carry the idea further and code P-values by their size. In this scheme, a P-value between 0.05 and 0.01 gets highlighted by \*. A P-value between 0.01 and 0.001 gets \*\*, and a P-value less than 0.001 gets \*\*\*. This can be a convenient summary of the weight of evidence against the null hypothesis if it's not taken too literally. But we warn you against taking the distinctions too seriously and against making a black-and-white decision near the boundaries. The boundaries are a matter of tradition, not science; there is nothing special about 0.05. A P-value of 0.051 should be looked at very seriously and not casually thrown away just because it's larger than 0.05, and one that's 0.009 is not very different from one that's 0.011.

When you decide to declare a verdict, it's always a good idea to report the P-value as an indication of the strength of the evidence. Sometimes it's best to report that the conclusion is not yet clear and to suggest that more data be gathered. (In a trial, a jury may “hang” and be unable to return a verdict.) In these cases, the P-value is the best summary we have of what the data say or fail to say about the null hypothesis.

## Significant vs. Important

### Practical vs. Statistical Significance

A large insurance company mined its data and found a statistically significant ( $P = 0.04$ ) difference between the mean value of policies sold in 2001 and 2002. The difference in the mean values was \$9.83. Even though it was statistically significant, management did not see this as an important difference when a typical policy sold for more than \$1000. On the other hand, even a clinically important improvement of 10% in cure rate with a new treatment is not likely to be statistically significant in a study of fewer than 225 patients. A small clinical trial would probably not be conclusive.

What do we mean when we say that a test is statistically significant? All we mean is that the test statistic had a P-value lower than our alpha level. Don't be lulled into thinking that statistical significance carries with it any sense of practical importance or impact.

For large samples, even small, unimportant (“insignificant”) deviations from the null hypothesis can be statistically significant. On the other hand, if the sample is not large enough, even large financially or scientifically “significant” differences may not be statistically significant.

It's good practice to report the magnitude of the difference between the observed statistic value and the null hypothesis value (in the data units) along with the P-value on which we base statistical significance.

## Confidence Intervals and Hypothesis Tests

For the motorcycle helmet example, a 95% confidence interval would give  $0.507 \pm 1.96 \times 0.0179 = (0.472, 0.542)$ , or 47.2% to 54.2%. If the previous rate of helmet compliance had been, say, 50%, we would not have been able to reject the null hypothesis because 50% is in the interval, so it's a plausible value. Indeed, *any* hypothesized value for the true proportion of helmet wearers in this interval is consistent with the data. Any value outside the confidence interval would make a null hypothesis that we would reject, but we'd feel more strongly about values far outside the interval.

Confidence intervals and hypothesis tests are built from the same calculations.<sup>5</sup> They have the same assumptions and conditions. As we have just seen, you can

<sup>5</sup> As we saw in Chapter 20, this is not *exactly* true for proportions. For a confidence interval, we estimate the standard deviation of  $\hat{p}$  from  $\hat{p}$  itself. Because we estimate it from the data, we have a *standard error*. For the corresponding hypothesis test, we use the model's standard deviation for  $\hat{p}$ , based on the null hypothesis value  $p_0$ . When  $\hat{p}$  and  $p_0$  are close, these calculations give similar results. When they differ, you're likely to reject  $H_0$  (because the observed proportion is far from your hypothesized value). In that case, you're better off building your confidence interval with a standard error estimated from the data.

approximate a hypothesis test by examining the confidence interval. Just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level. Because confidence intervals are naturally two-sided, they correspond to two-sided tests. For example, a 95% confidence interval corresponds to a two-sided hypothesis test at  $\alpha = 5\%$ . In general, a confidence interval with a confidence level of  $C\%$  corresponds to a two-sided hypothesis test with an  $\alpha$  level of  $100 - C\%$ .

The relationship between confidence intervals and one-sided hypothesis tests is a little more complicated. For a one-sided test with  $\alpha = 5\%$ , the corresponding confidence interval has a confidence level of 90%—that's 5% in each tail. In general, a confidence interval with a confidence level of  $C\%$  corresponds to a one-sided hypothesis test with an  $\alpha$  level of  $\frac{1}{2}(100 - C)\%$ .

**FOR EXAMPLE****Making a decision based on a confidence interval**

**Recap:** The baseline seven-year risk of heart attacks for diabetics is 20.2%. In 2007 a *NEJM* study reported a 95% confidence interval equivalent to 20.8% to 40.0% for the risk among patients taking the diabetes drug Avandia.

**Question:** What did this confidence interval suggest to the FDA about the safety of the drug?

The FDA could be 95% confident that the interval from 20.8% to 40.0% included the true risk of heart attack for diabetes patients taking Avandia. Because the lower limit of this interval was higher than the baseline risk of 20.2%, there was evidence of an increased risk.

**JUST CHECKING**

1. An experiment to test the fairness of a roulette wheel gives a z-score of 0.62. What would you conclude?
2. In the last chapter we encountered a bank that wondered if it could get more customers to make payments on delinquent balances by sending them a DVD urging them to set up a payment plan. Well, the bank just got back the results on their test of this strategy. A 90% confidence interval for the success rate is (0.29, 0.45). Their old send-a-letter method had worked 30% of the time. Can you reject the null hypothesis that the proportion is still 30% at  $\alpha = 0.05$ ? Explain.
3. Given the confidence interval the bank found in their trial of DVDs, what would you recommend that they do? Should they scrap the DVD strategy?

**STEP-BY-STEP EXAMPLE****Wear that Seatbelt!**

Teens are at the greatest risk of being killed or injured in traffic crashes. According to the National Highway Traffic Safety Administration, 65% of young people killed were not wearing a safety belt. In 2001, a total of 3322 teens were killed in motor vehicle crashes, an average of 9 teenagers a day. Because many of these deaths could easily be prevented by the use of safety belts, several states have begun "Click It or Ticket" campaigns in which increased enforcement and publicity have resulted in significantly higher seatbelt use. Overall use in Massachusetts quickly increased from 51% in 2002 to 64.8% in 2006, with a goal of surpassing the national average of 82%. Recently, a local newspaper reported that a roadblock resulted in 23 tickets to drivers who were unbelted out of 134 stopped for inspection.

**Question:** Does this provide evidence that the goal of over 82% compliance was met?

Let's use a confidence interval to test this hypothesis.

THINK

**Plan** State the problem and discuss the variables and the W's.

**Hypotheses** The null hypothesis is that the compliance rate is only 82%. The alternative is that it is now higher. It's clearly a one-sided test, so if we use a confidence interval, we'll have to be careful about what level we use.

**Model** Think about the assumptions and check the conditions.

We are finding a confidence interval, so we work from the data rather than the null model.

State your method.

The data come from a local newspaper report that tells the number of tickets issued and number of drivers stopped at a recent road-block. I want to know whether the rate of compliance with the seatbelt law is greater than 82%.

$$H_0: p = 0.82$$

$$H_A: p > 0.82$$

- ✓ **Independence Assumption:** Drivers are not likely to influence one another when it comes to wearing a seatbelt.
- ✓ **Randomization Condition:** This wasn't a random sample, but I assume these drivers are representative of the driving public.
- ✓ **10% Condition:** The police stopped fewer than 10% of all drivers.
- ✓ **Success/Failure Condition:** There were 111 successes and 23 failures, both at least 10. The sample is large enough.

Under these conditions, the sampling model is Normal. I'll create a **one-proportion z-interval**.

SHOW

**Mechanics** Write down the given information, and determine the sample proportion.

To use a confidence interval, we need a confidence level that corresponds to the alpha level of the test. If we use  $\alpha = 0.05$ , we should construct a 90% confidence interval, because this is a one-sided test.

That will leave 5% on *each* side of the observed proportion. Determine the standard error of the sample proportion and the margin of error. The critical value is  $z^* = 1.645$ .

The confidence interval is  
estimate  $\pm$  margin of error.

$n = 134$ , so

$$\hat{p} = \frac{111}{134} = 0.828 \text{ and}$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.828)(0.172)}{134}} = 0.033$$

$$ME = z^* \times SE(\hat{p}) \\ = 1.645(0.033) = 0.054$$

The 90% confidence interval is

$$0.828 \pm 0.054 \text{ or} \\ (0.774, 0.882).$$





**Conclusion** Link the confidence interval to your decision about the null hypothesis, and then state your conclusion in context.

I am 90% confident that between 77.4% and 88.2% of all drivers wear their seatbelts. Because the hypothesized rate of 82% is within this interval, I do not reject the null hypothesis. There is insufficient evidence to conclude that the campaign was truly effective and now more than 82% of all drivers are wearing seatbelts.

The upper limit of the confidence interval shows it's possible that the campaign is quite successful, but the small sample size makes the interval too wide to be very specific.

## \* A 95% Confidence Interval for Small Samples

When the **Success/Failure Condition** fails, all is not lost. A simple adjustment to the calculation lets us make a 95% confidence interval anyway.

All we do is add four *phony* observations—two to the successes, two to the failures. So instead of the proportion  $\hat{p} = \frac{y}{n}$ , we use the adjusted proportion  $\tilde{p} = \frac{y + 2}{n + 4}$  and, for convenience, we write  $\tilde{n} = n + 4$ . We modify the interval by using these adjusted values for both the center of the interval *and* the margin of error. Now the adjusted interval is

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}$$

This adjusted form gives better performance overall<sup>6</sup> and works much better for proportions near 0 or 1. It has the additional advantage that we no longer need to check the **Success/Failure Condition** that  $n\hat{p}$  and  $n\hat{q}$  are greater than 10.

### FOR EXAMPLE

#### An Agresti-Coull “plus-four” interval

Surgeons examined their results to compare two methods for a surgical procedure used to alleviate pain on the outside of the wrist. A new method was compared with the traditional “freehand” method for the procedure. Of 45 operations using the “freehand” method, three were unsuccessful, for a failure rate of 6.7%. With only 3 failures, the data don't satisfy the **Success/Failure Condition**, so we can't use a standard confidence interval.

**Question:** What's the confidence interval using the “plus-four” method?

<sup>6</sup> By “better performance,” we mean that a 95% confidence interval has more nearly a 95% chance of covering the true population proportion. Simulation studies have shown that our original, simpler confidence interval in fact is less likely than 95% to cover the true population proportion when the sample size is small or the proportion very close to 0 or 1. The original idea for this method can be attributed to E. B. Wilson. The simpler approach discussed here was proposed by Agresti and Coull (A. Agresti and B. A. Coull, “Approximate Is Better Than ‘Exact’ for Interval Estimation of Binomial Proportions,” *The American Statistician*, 52[1998]: 119–129).

There were 42 successes and 3 failures. Adding 2 “pseudo-successes” and 2 “pseudo-failures,” we find

$$\tilde{p} = \frac{3 + 2}{45 + 4} = 0.102$$

A 95% confidence interval is then

$$0.102 \pm 1.96 \sqrt{\frac{0.102(1 - 0.102)}{49}} = 0.102 \pm 0.085 \text{ or } (0.017, 0.187).$$

Notice that although the observed failure rate of 0.067 is contained in the interval, it is not at the center of the interval—something we haven't seen with any of the other confidence intervals we've considered.

## Making Errors

**AS** **Activity: Type I and Type II Errors.** View an animated exploration of Type I and Type II errors—a good backup for the reading in this section.

Some false-positive results mean no more than an unnecessary chest X-ray. But for a drug test or a disease like AIDS, a false-positive result that is not kept confidential could have serious consequences.

**AS** **Activity: Hypothesis Tests Are Random.** Simulate hypothesis tests and watch Type I errors occur. When you conduct real hypothesis tests you'll never know, but simulation can tell you when you've made an error.

Nobody's perfect. Even with lots of evidence, we can still make the wrong decision. In fact, when we perform a hypothesis test, we can make mistakes in *two* ways:

- I. The null hypothesis is true, but we mistakenly reject it.
- II. The null hypothesis is false, but we fail to reject it.

These two types of errors are known as **Type I** and **Type II errors**. One way to keep the names straight is to remember that we start by assuming the null hypothesis is true, so a Type I error is the first kind of error we could make.

In medical disease testing, the null hypothesis is usually the assumption that a person is healthy. The alternative is that he or she has the disease we're testing for. So a Type I error is a *false positive*: A healthy person is diagnosed with the disease. A Type II error, in which an infected person is diagnosed as disease free, is a *false negative*. These errors have other names, depending on the particular discipline and context.

Which type of error is more serious depends on the situation. In the jury trial, a Type I error occurs if the jury convicts an innocent person. A Type II error occurs if the jury fails to convict a guilty person. Which seems more serious? In medical diagnosis, a false negative could mean that a sick patient goes untreated. A false positive might mean that the person must undergo further tests. In a Statistics final exam (with  $H_0$ : the student has learned only 60% of the material), a Type I error would be passing a student who in fact learned less than 60% of the material, while a Type II error would be failing a student who knew enough to pass. Which of these errors seems more serious? It depends on the situation, the cost, and your point of view.

Here's an illustration of the situations:

		The Truth	
		$H_0$ True	$H_0$ False
My Decision	Reject $H_0$	Type I Error	OK
	Fail to reject $H_0$	OK	Type II Error

How often will a Type I error occur? It happens when the null hypothesis is true but we've had the bad luck to draw an unusual sample. To reject  $H_0$ , the P-value

**NOTATION ALERT:**

In Statistics,  $\alpha$  is almost always saved for the alpha level. But  $\beta$  has already been used for the parameters of a linear model. Fortunately, it's usually clear whether we're talking about a Type II error probability or the slope or intercept of a regression model.

The null hypothesis specifies a single value for the parameter. So it's easy to calculate the probability of a Type I error. But the alternative gives a whole range of possible values, and we may want to find a  $\beta$  for several of them.

We have seen ways to find a sample size by specifying the margin of error. Choosing the sample size to achieve a specified  $\beta$  (for a particular alternative value) is sometimes more appropriate, but the calculation is more complex and lies beyond the scope of this book.

must fall below  $\alpha$ . When  $H_0$  is true, that happens *exactly* with probability  $\alpha$ . So when you choose level  $\alpha$ , you're setting the probability of a Type I error to  $\alpha$ .

What if  $H_0$  is not true? Then we can't possibly make a Type I error. You can't get a false positive from a sick person. A Type I error can happen only when  $H_0$  is true.

When  $H_0$  is false and we reject it, we have done the right thing. A test's ability to detect a false null hypothesis is called the **power** of the test. In a jury trial, power is the ability of the criminal justice system to convict people who are guilty—a good thing! We'll have a lot more to say about power soon.

When  $H_0$  is false but we fail to reject it, we have made a Type II error. We assign the letter  $\beta$  to the probability of this mistake. What's the value of  $\beta$ ? That's harder to assess than  $\alpha$  because we don't know what the value of the parameter really is. When  $H_0$  is true, it specifies a single parameter value. But when  $H_0$  is false, we don't have a specific one; we have many possible values. We can compute the probability  $\beta$  for any parameter value in  $H_A$ . But which one should we choose?

One way to focus our attention is by thinking about the *effect size*. That is, we ask "How big a difference would matter?" Suppose a charity wants to test whether placing personalized address labels in the envelope along with a request for a donation increases the response rate above the baseline of 5%. If the minimum response that would pay for the address labels is 6%, they would calculate  $\beta$  for the alternative  $p = 0.06$ .

We could reduce  $\beta$  for *all* alternative parameter values by increasing  $\alpha$ . By making it easier to reject the null, we'd be more likely to reject it whether it's true or not. So we'd reduce  $\beta$ , the chance that we fail to reject a false null—but we'd make more Type I errors. This tension between Type I and Type II errors is inevitable. In the political arena, think of the ongoing debate between those who favor provisions to reduce Type I errors in the courts (supporting Miranda rights, requiring warrants for wiretaps, providing legal representation for those who can't afford it) and those who advocate changes to reduce Type II errors (admitting into evidence confessions made when no lawyer is present, eavesdropping on conferences with lawyers, restricting paths of appeal, etc.).

The only way to reduce *both* types of error is to collect more evidence or, in statistical terms, to collect more data. Too often, studies fail because their sample sizes are too small to detect the change they are looking for.

**FOR EXAMPLE****Thinking about errors**

**Recap:** A published study found the risk of heart attack to be increased in patients taking the diabetes drug Avandia. The issue of the *New England Journal of Medicine (NEJM)* in which that study appeared also included an editorial that said, in part, "A few events either way might have changed the findings for myocardial infarction<sup>7</sup> or for death from cardiovascular causes. In this setting, the possibility that the findings were due to chance cannot be excluded."

**Question:** What kind of error would the researchers have made if, in fact, their findings were due to chance? What could be the consequences of this error?

The null hypothesis said the risk didn't change, but the researchers rejected that model and claimed evidence of a higher risk. If these findings were just due to chance, they rejected a true null hypothesis—a Type I error.

If, in fact, Avandia carried no extra risk, then patients might be deprived of its benefits for no good reason.

<sup>7</sup> Doctorese for "heart attack."

## Power

When we failed to reject the null hypothesis about TT practitioners, did we prove that they were just guessing? No, it could be that they actually *can* discern a human energy field but we just couldn't tell. For example, suppose they really have the ability to get 53% of the trials right but just happened to get only 47% in our experiment. Our confidence interval shows that with these data we wouldn't have rejected the null. And if we retained the null even though the true proportion was actually greater than 50%, we would have made a Type II error because we failed to detect their ability.

Remember, we can never prove a null hypothesis true. We can only fail to reject it. But when we fail to reject a null hypothesis, it's natural to wonder whether we looked hard enough. Might the null hypothesis actually be false and our test too weak to tell?

When the null hypothesis actually *is* false, we hope our test is strong enough to reject it. We'd like to know how likely we are to succeed. The power of the test gives us a way to think about that. **The power of a test is the probability that it correctly rejects a false null hypothesis.** When the power is high, we can be confident that we've looked hard enough. We know that  $\beta$  is the probability that a test *fails* to reject a false null hypothesis, so the power of the test is the probability that it *does* reject:  $1 - \beta$ .

Whenever a study fails to reject its null hypothesis, the test's power comes into question. Was the sample size big enough to detect an effect had there been one? Might we have missed an effect large enough to be interesting just because we failed to gather sufficient data or because there was too much variability in the data we could gather? The therapeutic touch experiment failed to reject the null hypothesis that the TT practitioners were just guessing. Might the problem be that the experiment simply lacked adequate power to detect their ability?

### FOR EXAMPLE

#### Errors and power

**Recap:** The study of Avandia published in the *NEJM* combined results from 47 different trials—a method called *meta-analysis*. The drug's manufacturer, GlaxoSmithKline (GSK), issued a statement that pointed out, "Each study is designed differently and looks at unique questions: For example, individual studies vary in size and length, in the type of patients who participated, and in the outcomes they investigate." Nevertheless, by combining data from many studies, meta-analyses can achieve a much larger sample size.

**Question:** How could this larger sample size help?

If Avandia really did increase the seven-year heart attack rate, doctors needed to know. To overlook that would have been a Type II error (failing to detect a false null hypothesis), resulting in patients being put at greater risk. Increasing the sample size could increase the power of the analysis, making it more likely that researchers will detect the danger if there is one.

### AS

**Activity: The Power of a Test.** Power is a concept that's much easier to understand when you can visualize what's happening.

When we calculate power, we imagine that the null hypothesis is false. The value of the power depends on how far the truth lies from the null hypothesis value. **We call the distance between the null hypothesis value,  $p_0$ , and the truth,  $p$ , the effect size.** The power depends directly on the effect size. It's easier to see larger effects, so the farther  $p_0$  is from  $p$ , the greater the power. If the therapeutic touch practitioners were in fact able to detect human energy fields 90% of the time, it should be easy to see that they aren't guessing. With an effect size this large, we'd have a powerful test. If their true success rate were only 53%, however, we'd need a larger sample size to have a good chance of noticing that (and rejecting  $H_0$ ).

How can we decide what power we need? Choice of power is more a financial or scientific decision than a statistical one because to calculate the power, we need to specify the "true" parameter value we're interested in. In other words,

power is calculated for a particular effect size, and it changes depending on the size of the effect we want to detect. For example, do you think that health insurance companies should pay for therapeutic touch if practitioners could detect a human energy field only 53% of the time—just slightly better than chance? That doesn't seem clinically useful.<sup>8</sup> How about 75% of the time? No therapy works all the time, and insurers might be quite willing to pay for such a success rate. Let's take 75% as a reasonably interesting effect size (keeping in mind that 50% is the level of guessing). With 150 trials, the TT experiment would have been able to detect such an ability with a power of 99.99%. So power was not an issue in this study. There is only a very small chance that the study would have failed to detect a practitioner's ability, had it existed. The sample size was clearly big enough.



### JUST CHECKING

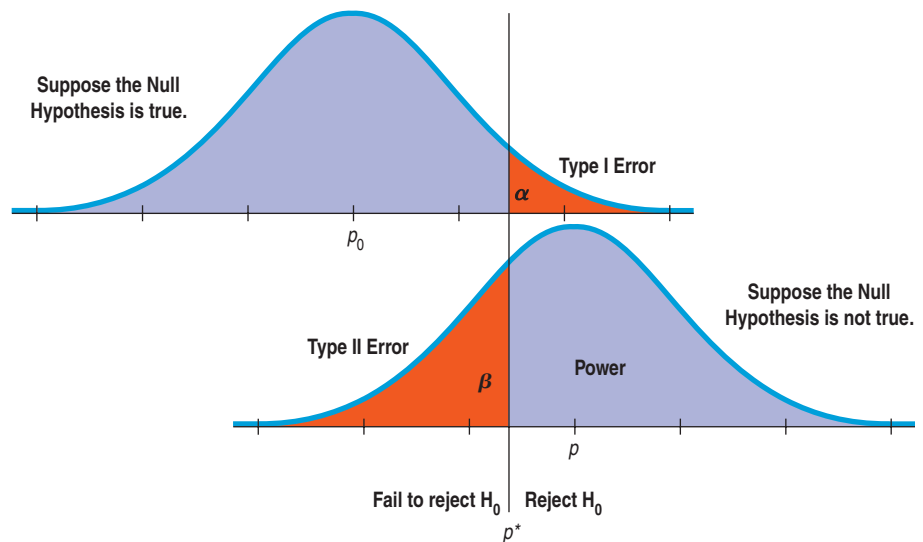
- Remember our bank that's sending out DVDs to try to get customers to make payments on delinquent loans? It is looking for evidence that the costlier DVD strategy produces a higher success rate than the letters it has been sending. Explain what a Type I error is in this context and what the consequences would be to the bank.
- What's a Type II error in the bank experiment context, and what would the consequences be?
- For the bank, which situation has higher power: a strategy that works really well, actually getting 60% of people to pay off their balances, or a strategy that barely increases the payoff rate to 32%? Explain briefly.

## A Picture Worth $\frac{1}{P(z > 3.09)}$ Words

It makes intuitive sense that the larger the effect size, the easier it should be to see it. Obtaining a larger sample size decreases the probability of a Type II error, so it increases the power. It also makes sense that the more we're willing to accept a Type I error, the less likely we will be to make a Type II error.

FIGURE 21.1

The power of a test is the probability that it rejects a false null hypothesis. The upper figure shows the null hypothesis model. We'd reject the null in a one-sided test if we observed a value of  $\hat{p}$  in the red region to the right of the critical value,  $p^*$ . The lower figure shows the true model. If the true value of  $p$  is greater than  $p_0$ , then we're more likely to observe a value that exceeds the critical value and make the correct decision to reject the null hypothesis. The power of the test is the purple region on the right of the lower figure. Of course, even drawing samples whose observed proportions are distributed around  $p$ , we'll sometimes get a value in the red region on the left and make a Type II error of failing to reject the null.



<sup>8</sup> On the other hand, a scientist might be interested in anything clearly different from the 50% guessing rate because that might suggest an entirely new physics at work. In fact, it could lead to a Nobel prize.

**NOTATION ALERT:**

We've attached symbols to many of the  $p$ 's. Let's keep them straight.  $p$  is a true proportion parameter.  $p_0$  is a hypothesized value of  $p$ .  $\hat{p}$  is an observed proportion.  $p^*$  is a critical value of a proportion corresponding to a specified  $\alpha$ .

**Fisher and  $\alpha = 0.05$** 

Why did Sir Ronald Fisher suggest 0.05 as a criterion for testing hypotheses? It turns out that he had in mind small initial studies. Small studies have relatively little power. Fisher was concerned that they might make too many Type II errors—failing to discover an important effect—if too strict a criterion were used. Once a test failed to reject a null hypothesis, it was unlikely that researchers would return to that hypothesis to try again.

On the other hand, the increased risk of Type I errors arising from a generous criterion didn't concern him as much for exploratory studies because these are ordinarily followed by a replication or a larger study. The probability of a Type I error is  $\alpha$ —in this case, 0.05. The probability that two independent studies would both make Type I errors is  $0.05 \times 0.05 = 0.0025$ , so Fisher was confident that Type I errors in initial studies were not a major concern.

The widespread use of the relatively generous 0.05 criterion even in large studies is most likely not what Fisher had in mind.

Figure 21.1 shows a good way to visualize the relationships among these concepts. Suppose we are testing  $H_0: p = p_0$  against the alternative  $H_A: p > p_0$ . We'll reject the null if the observed proportion,  $\hat{p}$ , is big enough. By big enough, we mean  $\hat{p} > p^*$  for some critical value,  $p^*$  (shown as the red region in the right tail of the upper curve). For example, we might be willing to believe the ability of therapeutic touch practitioners if they were successful in 65% of our trials. This is what the upper model shows. It's a picture of the sampling distribution model for the proportion if the null hypothesis were true. We'd make a Type I error whenever the sample gave us  $\hat{p} > p^*$ , because we would reject the (true) null hypothesis. And unusual samples like that would happen only with probability  $\alpha$ .

In reality, though, the null hypothesis is rarely *exactly* true. The lower probability model supposes that  $H_0$  is not true. In particular, it supposes that the true value is  $p$ , not  $p_0$ . (Perhaps the TT practitioner really can detect the human energy field 72% of the time.) It shows a distribution of possible observed  $\hat{p}$  values around this true value. Because of sampling variability, sometimes  $\hat{p} < p^*$  and we fail to reject the (false) null hypothesis. Suppose a TT practitioner with a true ability level of 72% is actually successful on fewer than 65% of our tests. Then we'd make a Type II error. The area under the curve to the left of  $p^*$  in the bottom model represents how often this happens. The probability is  $\beta$ . In this picture,  $\beta$  is less than half, so most of the time we *do* make the right decision. The *power* of the test—the probability that we make the right decision—is shown as the region to the right of  $p^*$ . It's  $1 - \beta$ .

We calculate  $p^*$  based on the upper model because  $p^*$  depends only on the null model and the alpha level. No matter what the true proportion, no matter whether the practitioners can detect a human energy field 90%, 53%, or 2% of the time,  $p^*$  doesn't change. After all, we don't *know* the truth, so we can't use it to determine the critical value. But we always reject  $H_0$  when  $\hat{p} > p^*$ .

How often we correctly reject  $H_0$  when it's *false* depends on the effect size. We can see from the picture that if the effect size were larger (the true proportion were farther above the hypothesized value), the bottom curve would shift to the right, making the power greater.

We can see several important relationships from this figure:

- ▶ Power =  $1 - \beta$ .
- ▶ Reducing  $\alpha$  to lower the chance of committing a Type I error will move the critical value,  $p^*$ , to the right (in this example). This will have the effect of increasing  $\beta$ , the probability of a Type II error, and correspondingly reducing the power.
- ▶ The larger the real difference between the hypothesized value,  $p_0$ , and the true population value,  $p$ , the smaller the chance of making a Type II error and the greater the power of the test. If the two proportions are very far apart, the two models will barely overlap, and we will not be likely to make any Type II errors at all—but then, we are unlikely to really need a formal hypothesis-testing procedure to see such an obvious difference. If the TT practitioners were successful almost all the time, we'd be able to see that with even a small experiment.

## Reducing Both Type I and Type II Errors

**A S**

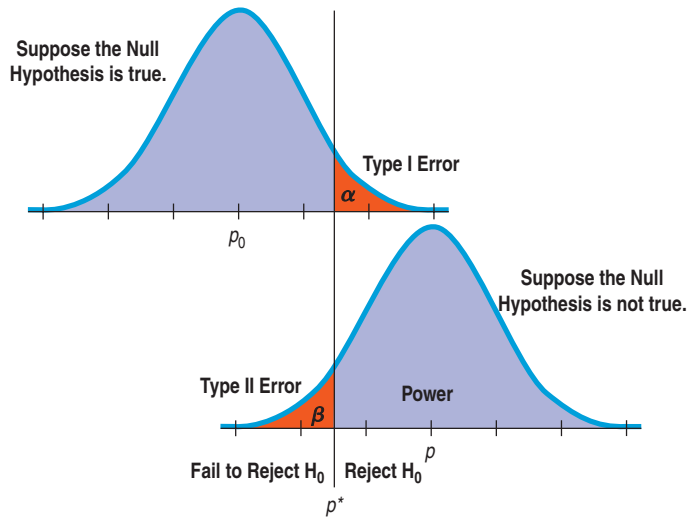
**Activity: Power and Sample Size.** Investigate how the power of a test changes with the sample size. The interactive tool is really the only way you can see this easily.

Figure 21.1 seems to show that if we reduce Type I error, we automatically must increase Type II error. But there is a way to reduce both. Can you think of it?

If we can make both curves narrower, as shown in Figure 21.2, then both the probability of Type I errors and the probability of Type II errors will decrease, and the power of the test will increase.

FIGURE 21.2

Making the standard deviations smaller increases the power without changing the corresponding critical value. The means are just as far apart as in Figure 21.1, but the error rates are reduced.

TI-*inspire*

**Errors and power.** Explore the relationships among Type I and Type II errors, sample size, effect size, and the power of a test.

How can we accomplish that? The only way is to reduce the standard deviations by increasing the sample size. (Remember, these are pictures of sampling distribution models, not of data.) Increasing the sample size works regardless of the true population parameters. But recall the curse of diminishing returns. The standard deviation of the sampling distribution model decreases only as the *square root* of the sample size, so to halve the standard deviations we must *quadruple* the sample size.

## FOR EXAMPLE

## Sample size, errors, and power

**Recap:** The meta-analysis of the risks of heart attacks in patients taking the diabetes drug Avandia combined results from 47 smaller studies. As GlaxoSmith-Kline (GSK), the drug's manufacturer, pointed out in their rebuttal, "Data from the ADOPT clinical trial did show a small increase in reports of myocardial infarction among the *Avandia*-treated group . . . however, the number of events is too small to reach a reliable conclusion about the role any of the medicines may have played in this finding."

**Question:** Why would this smaller study have been less likely to detect the difference in risk? What are the appropriate statistical concepts for comparing the smaller studies?

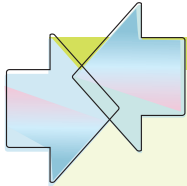
Smaller studies are subject to greater sampling variability; that is, the sampling distributions they estimate have a larger standard deviation for the sample proportion. That gives small studies less power: They'd be less able to discern whether an apparently higher risk was merely the result of chance variation or evidence of real danger. The FDA doesn't want to restrict the use of a drug that's safe and effective (Type I error), nor do they want patients to continue taking a medication that puts them at risk (Type II error). Larger sample sizes can reduce the risk of both kinds of error. Greater power (the probability of rejecting a false null hypothesis) means a better chance of spotting a genuinely higher risk of heart attacks.

## WHAT CAN GO WRONG?

- ▶ **Don't interpret the P-value as the probability that  $H_0$  is true.** The P-value is about the data, not the hypothesis. It's the probability of observing data this unusual, *given* that  $H_0$  is true, not the other way around.
- ▶ **Don't believe too strongly in arbitrary alpha levels.** There's not really much difference between a P-value of 0.051 and a P-value of 0.049, but sometimes it's regarded as the difference between night (having to refrain from rejecting  $H_0$ ) and day (being able to

shout to the world that your results are “statistically significant”). It may just be better to report the P-value and a confidence interval and let the world decide along with you.

- ▶ **Don't confuse practical and statistical significance.** A large sample size can make it easy to discern even a trivial change from the null hypothesis value. On the other hand, an important difference can be missed if your test lacks sufficient power.
- ▶ **Don't forget that in spite of all your care, you might make a wrong decision.** We can never reduce the probability of a Type I error ( $\alpha$ ) or of a Type II error ( $\beta$ ) to zero (but increasing the sample size helps).



## CONNECTIONS

All of the hypothesis tests we'll see boil down to the same question: “Is the difference between two quantities large?” We always measure “how large” by finding a ratio of this difference to the standard deviation of the sampling distribution of the statistic. Using the standard deviation as our ruler for inference is one of the core ideas of statistical thinking.

We've discussed the close relationship between hypothesis tests and confidence intervals. They are two sides of the same coin.

This chapter also has natural links to the discussion of probability, to the Normal model, and to the two previous chapters on inference.

## WHAT HAVE WE LEARNED?



We've learned that there's a lot more to hypothesis testing than a simple yes/no decision.

- ▶ We've learned that the P-value can indicate evidence against the null hypothesis when it's small, but it does not tell us the probability that the null hypothesis is true.
- ▶ We've learned that the alpha level of the test establishes the level of proof we'll require. That determines the critical value of  $z$  that will lead us to reject the null hypothesis.
- ▶ We've also learned more about the connection between hypothesis tests and confidence intervals; they're really two ways of looking at the same question. The hypothesis test gives us the answer to a decision about a parameter; the confidence interval tells us the plausible values of that parameter.

We've learned about the two kinds of errors we might make, and we've seen why in the end we're never sure we've made the right decision.

- ▶ If the null hypothesis is really true and we reject it, that's a Type I error; the alpha level of the test is the probability that this could happen.
- ▶ If the null hypothesis is really false but we fail to reject it, that's a Type II error.
- ▶ The power of the test is the probability that we reject the null hypothesis when it's false. The larger the size of the effect we're testing for, the greater the power of the test to detect it.
- ▶ We've seen that tests with a greater likelihood of Type I error have more power and less chance of a Type II error. We can increase power while reducing the chances of both kinds of error by increasing the sample size.

## Terms

Alpha level

486. The threshold P-value that determines when we reject a null hypothesis. If we observe a statistic whose P-value based on the null hypothesis is less than  $\alpha$ , we reject that null hypothesis.

Statistically significant

486. When the P-value falls below the alpha level, we say that the test is “statistically significant” at that alpha level.



<b>Significance level</b>	486. The alpha level is also called the significance level, most often in a phrase such as a conclusion that a particular test is “significant at the 5% significance level.”
<b>Type I error</b>	491. The error of rejecting a null hypothesis when in fact it is true (also called a “false positive”). The probability of a Type I error is $\alpha$ .
<b>Type II error</b>	491. The error of failing to reject a null hypothesis when in fact it is false (also called a “false negative”). The probability of a Type II error is commonly denoted $\beta$ and depends on the effect size.
<b>Power</b>	492, 493. The probability that a hypothesis test will correctly reject a false null hypothesis is the power of the test. To find power, we must specify a particular alternative parameter value as the “true” value. For any specific value in the alternative, the power is $1 - \beta$ .
<b>Effect size</b>	493. The difference between the null hypothesis value and true value of a model parameter is called the effect size.

## Skills

### THINK

- ▶ Understand that statistical significance does not measure the importance or magnitude of an effect. Recognize when others misinterpret statistical significance as proof of practical importance.
- ▶ Understand the close relationship between hypothesis tests and confidence intervals.
- ▶ Be able to identify and use the alternative hypothesis when testing hypotheses. Understand how to choose between a one-sided and two-sided alternative hypothesis, and know how to defend the choice of a one-sided alternative.
- ▶ Understand how the critical value for a test is related to the specified alpha level.
- ▶ Understand that the power of a test gives the probability that it correctly rejects a false null hypothesis when a specified alternative is true.
- ▶ Understand that the power of a test depends in part on the sample size. Larger sample sizes lead to greater power (and thus fewer Type II errors).

### SHOW

- ▶ Know how to complete a hypothesis test for a population proportion.

### TELL

- ▶ Be able to interpret the meaning of a P-value in nontechnical language.
- ▶ Understand that the P-value of a test does not give the probability that the null hypothesis is correct.
- ▶ Know that we do not “accept” a null hypothesis if we cannot reject it but, rather, that we can only “fail to reject” the hypothesis for lack of evidence against it.

## HYPOTHESIS TESTS ON THE COMPUTER

Reports about hypothesis tests generated by technologies don't follow a standard form. Most will name the test and provide the test statistic value, its standard deviation, and the P-value. But these elements may not be labeled clearly. For example, the expression “Prob > |z|” means the probability (the “Prob”) of observing a test statistic whose magnitude (the absolute value tells us this) is larger than that of the one (the “z”) found in the data (which, because it is written as “z,” we know follows a Normal model). That is a fancy (and not very clear) way of saying P-value. In some packages, you can specify that the test be one-sided. Others might report three P-values, covering the ground for both one-sided tests and the two-sided test.

Sometimes a confidence interval and hypothesis test are automatically given together. The CI ought to be for the corresponding confidence level:  $1 - \alpha$  for 2-tailed tests,  $1 - 2\alpha$  for 1-tailed tests.

Often, the standard deviation of the statistic is called the “standard error,” and usually that's appropriate because we've had to estimate its value from the data. That's not the case for proportions, however: We get the

standard deviation for a proportion from the null hypothesis value. Nevertheless, you may see the standard deviation called a “standard error” even for tests with proportions.

It’s common for statistics packages and calculators to report more digits of “precision” than could possibly have been found from the data. You can safely ignore them. Round values such as the standard deviation to one digit more than the number of digits reported in your data.

Here are the kind of results you might see. This is not from any program or calculator we know of, but it shows some of the things you might see in typical computer output.

usually, the test is named

	Value	Test Stat	Prob >  Z
Estimate	0.467	-0.825	0.42
Std Err	0.04073		
Upper 95%	0.547		
Lower 95%	0.387		

Actually, a standard deviation because this is a test

Might offer a CI as well  
These are bounds for the 95% CI because  $\alpha = 0.05$ —a fact not clearly stated

test statistic value

P-value

2-sided alternative

For information on hypothesis testing with particular statistics packages, see the table for Chapter 20 in Appendix B.

## EXERCISES

- One sided or two?** In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?
  - A business student conducts a taste test to see whether students prefer Diet Coke or Diet Pepsi.
  - PepsiCo recently reformulated Diet Pepsi in an attempt to appeal to teenagers. They run a taste test to see if the new formula appeals to more teenagers than the standard formula.
  - A budget override in a small town requires a two-thirds majority to pass. A local newspaper conducts a poll to see if there’s evidence it will pass.
  - One financial theory states that the stock market will go up or down with equal probability. A student collects data over several years to test the theory.
- Which alternative?** In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?
  - A college dining service conducts a survey to see if students prefer plastic or metal cutlery.
  - In recent years, 10% of college juniors have applied for study abroad. The dean’s office conducts a survey to see if that’s changed this year.
  - A pharmaceutical company conducts a clinical trial to see if more patients who take a new drug experience headache relief than the 22% who claimed relief after taking the placebo.
  - At a small computer peripherals company, only 60% of the hard drives produced passed all their performance tests the first time. Management recently invested a lot of resources into the production system and now conducts a test to see if it helped.
- P-value.** A medical researcher tested a new treatment for poison ivy against the traditional ointment. He concluded that the new treatment is more effective. Explain what the P-value of 0.047 means in this context.
- Another P-value.** Have harsher penalties and ad campaigns increased seat-belt use among drivers and passengers? Observations of commuter traffic failed to find evidence of a significant change compared with three years ago. Explain what the study’s P-value of 0.17 means in this context.
- Alpha.** A researcher developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. He

standard deviation for a proportion from the null hypothesis value. Nevertheless, you may see the standard deviation called a “standard error” even for tests with proportions.

It’s common for statistics packages and calculators to report more digits of “precision” than could possibly have been found from the data. You can safely ignore them. Round values such as the standard deviation to one digit more than the number of digits reported in your data.

Here are the kind of results you might see. This is not from any program or calculator we know of, but it shows some of the things you might see in typical computer output.

usually, the test is named

	Value	Test Stat	Prob >  Z
Estimate	0.467	-0.825	0.42
Std Err	0.04073		
Upper 95%	0.547		
Lower 95%	0.387		

Actually, a standard deviation because this is a test

Might offer a CI as well  
These are bounds for the 95% CI because  $\alpha = 0.05$ —a fact not clearly stated

test statistic value

P-value

2-sided alternative

For information on hypothesis testing with particular statistics packages, see the table for Chapter 20 in Appendix B.

## EXERCISES

- One sided or two?** In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?
  - A business student conducts a taste test to see whether students prefer Diet Coke or Diet Pepsi.
  - PepsiCo recently reformulated Diet Pepsi in an attempt to appeal to teenagers. They run a taste test to see if the new formula appeals to more teenagers than the standard formula.
  - A budget override in a small town requires a two-thirds majority to pass. A local newspaper conducts a poll to see if there’s evidence it will pass.
  - One financial theory states that the stock market will go up or down with equal probability. A student collects data over several years to test the theory.
- Which alternative?** In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?
  - A college dining service conducts a survey to see if students prefer plastic or metal cutlery.
  - In recent years, 10% of college juniors have applied for study abroad. The dean’s office conducts a survey to see if that’s changed this year.
  - A pharmaceutical company conducts a clinical trial to see if more patients who take a new drug experience headache relief than the 22% who claimed relief after taking the placebo.
  - At a small computer peripherals company, only 60% of the hard drives produced passed all their performance tests the first time. Management recently invested a lot of resources into the production system and now conducts a test to see if it helped.
- P-value.** A medical researcher tested a new treatment for poison ivy against the traditional ointment. He concluded that the new treatment is more effective. Explain what the P-value of 0.047 means in this context.
- Another P-value.** Have harsher penalties and ad campaigns increased seat-belt use among drivers and passengers? Observations of commuter traffic failed to find evidence of a significant change compared with three years ago. Explain what the study’s P-value of 0.17 means in this context.
- Alpha.** A researcher developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. He

- made this decision based on a test using  $\alpha = 0.05$ . Would he have made the same decision at  $\alpha = 0.10$ ? How about  $\alpha = 0.01$ ? Explain.
- Alpha again.** Environmentalists concerned about the impact of high-frequency radio transmissions on birds found that there was no evidence of a higher mortality rate among hatchlings in nests near cell towers. They based this conclusion on a test using  $\alpha = 0.05$ . Would they have made the same decision at  $\alpha = 0.10$ ? How about  $\alpha = 0.01$ ? Explain.
  - Significant?** Public health officials believe that 90% of children have been vaccinated against measles. A random survey of medical records at many schools across the country found that, among more than 13,000 children, only 89.4% had been vaccinated. A statistician would reject the 90% hypothesis with a P-value of  $P = 0.011$ .
    - Explain what the P-value means in this context.
    - The result is statistically significant, but is it important? Comment.
  - Significant again?** A new reading program may reduce the number of elementary school students who read below grade level. The company that developed this program supplied materials and teacher training for a large-scale test involving nearly 8500 children in several different school districts. Statistical analysis of the results showed that the percentage of students who did not meet the grade-level goal was reduced from 15.9% to 15.1%. The hypothesis that the new reading program produced no improvement was rejected with a P-value of 0.023.
    - Explain what the P-value means in this context.
    - Even though this reading method has been shown to be significantly better, why might you not recommend that your local school adopt it?
  - Success.** In August 2004, *Time* magazine reported the results of a random telephone poll commissioned by the Spike network. Of the 1302 men who responded, only 39 said that their most important measure of success was their work.
    - Estimate the percentage of all American males who measure success primarily from their work. Use a 98% confidence interval. Check the conditions first.
    - Some believe that few contemporary men judge their success primarily by their work. Suppose we wished to conduct a hypothesis test to see if the fraction has fallen below the 5% mark. What does your confidence interval indicate? Explain.
    - What is the level of significance of this test? Explain.
  - Is the Euro fair?** Soon after the Euro was introduced as currency in Europe, it was widely reported that someone had spun a Euro coin 250 times and gotten heads 140 times. We wish to test a hypothesis about the fairness of spinning the coin.
    - Estimate the true proportion of heads. Use a 95% confidence interval. Don't forget to check the conditions.
    - Does your confidence interval provide evidence that the coin is unfair when spun? Explain.
    - What is the significance level of this test? Explain.
  - Approval 2007.** In May 2007, George W. Bush's approval rating stood at 30% according to a CBS News/*New York Times* national survey of 1125 randomly selected adults.
    - Make a 95% confidence interval for his approval rating by all U.S. adults.
    - Based on the confidence interval, test the null hypothesis that Bush's approval rating was no better than the 27% level established by Richard Nixon during the Watergate scandal.
  - Superdads.** The Spike network commissioned a telephone poll of randomly sampled U.S. men. Of the 712 respondents who had children, 22% said "yes" to the question "Are you a stay-at-home dad?" (*Time*, August 23, 2004)
    - To help market commercial time, Spike wants an accurate estimate of the true percentage of stay-at-home dads. Construct a 95% confidence interval.
    - An advertiser of baby-carrying slings for dads will buy commercial time if at least 25% of men are stay-at-home dads. Use your confidence interval to test an appropriate hypothesis, and make a recommendation to the advertiser.
    - Could Spike claim to the advertiser that it is possible that 25% of men with young children are stay-at-home dads? What is wrong with the reasoning?
  - Dogs.** Canine hip dysplasia is a degenerative disease that causes pain in many dogs. Sometimes advanced warning signs appear in puppies as young as 6 months. A veterinarian checked 42 puppies whose owners brought them to a vaccination clinic, and she found 5 with early hip dysplasia. She considers this group to be a random sample of all puppies.
    - Explain we cannot use this information to construct a confidence interval for the rate of occurrence of early hip dysplasia among all 6-month-old puppies.
    - \*b) Construct a "plus-four" confidence interval and interpret it in this context.
  - Fans.** A survey of 81 randomly selected people standing in line to enter a football game found that 73 of them were home team fans.
    - Explain why we cannot use this information to construct a confidence interval for the proportion of all people at the game who are fans of the home team.
    - \*b) Construct a "plus-four" confidence interval and interpret it in this context.
  - Loans.** Before lending someone money, banks must decide whether they believe the applicant will repay the loan. One strategy used is a point system. Loan officers assess information about the applicant, totaling points they award for the person's income level, credit history, current debt burden, and so on. The higher the point total, the more convinced the bank is that it's safe to make the loan. Any applicant with a lower point total than a certain cutoff score is denied a loan.
 

We can think of this decision as a hypothesis test. Since the bank makes its profit from the interest collected on repaid loans, their null hypothesis is that the applicant will repay the loan and therefore should get the money. Only if the person's score falls below the minimum cutoff will the bank reject the null and deny

the loan. This system is reasonably reliable, but, of course, sometimes there are mistakes.

- a) When a person defaults on a loan, which type of error did the bank make?
  - b) Which kind of error is it when the bank misses an opportunity to make a loan to someone who would have repaid it?
  - c) Suppose the bank decides to lower the cutoff score from 250 points to 200. Is that analogous to choosing a higher or lower value of  $\alpha$  for a hypothesis test? Explain.
  - d) What impact does this change in the cutoff value have on the chance of each type of error?
- 16. Spam.** Spam filters try to sort your e-mails, deciding which are real messages and which are unwanted. One method used is a point system. The filter reads each incoming e-mail and assigns points to the sender, the subject, key words in the message, and so on. The higher the point total, the more likely it is that the message is unwanted. The filter has a cutoff value for the point total; any message rated lower than that cutoff passes through to your inbox, and the rest, suspected to be spam, are diverted to the junk mailbox.
- We can think of the filter's decision as a hypothesis test. The null hypothesis is that the e-mail is a real message and should go to your inbox. A higher point total provides evidence that the message may be spam; when there's sufficient evidence, the filter rejects the null, classifying the message as junk. This usually works pretty well, but, of course, sometimes the filter makes a mistake.
- a) When the filter allows spam to slip through into your inbox, which kind of error is that?
  - b) Which kind of error is it when a real message gets classified as junk?
  - c) Some filters allow the user (that's you) to adjust the cutoff. Suppose your filter has a default cutoff of 50 points, but you reset it to 60. Is that analogous to choosing a higher or lower value of  $\alpha$  for a hypothesis test? Explain.
  - d) What impact does this change in the cutoff value have on the chance of each type of error?
- 17. Second loan.** Exercise 15 describes the loan score method a bank uses to decide which applicants it will lend money. Only if the total points awarded for various aspects of an applicant's financial condition fail to add up to a minimum cutoff score set by the bank will the loan be denied.
- a) In this context, what is meant by the power of the test?
  - b) What could the bank do to increase the power?
  - c) What's the disadvantage of doing that?
- 18. More spam.** Consider again the points-based spam filter described in Exercise 16. When the points assigned to various components of an e-mail exceed the cutoff value you've set, the filter rejects its null hypothesis (that the message is real) and diverts that e-mail to a junk mailbox.
- a) In this context, what is meant by the power of the test?
  - b) What could you do to increase the filter's power?
  - c) What's the disadvantage of doing that?
- 19. Homeowners 2005.** In 2005 the U.S. Census Bureau reported that 68.9% of American families owned their homes. Census data reveal that the ownership rate in one small city is much lower. The city council is debating a plan to offer tax breaks to first-time home buyers in order to encourage people to become homeowners. They decide to adopt the plan on a 2-year trial basis and use the data they collect to make a decision about continuing the tax breaks. Since this plan costs the city tax revenues, they will continue to use it only if there is strong evidence that the rate of home ownership is increasing.
- a) In words, what will their hypotheses be?
  - b) What would a Type I error be?
  - c) What would a Type II error be?
  - d) For each type of error, tell who would be harmed.
  - e) What would the power of the test represent in this context?
- 20. Alzheimer's.** Testing for Alzheimer's disease can be a long and expensive process, consisting of lengthy tests and medical diagnosis. Recently, a group of researchers (Solomon *et al.*, 1998) devised a 7-minute test to serve as a quick screen for the disease for use in the general population of senior citizens. A patient who tested positive would then go through the more expensive battery of tests and medical diagnosis. The authors reported a false positive rate of 4% and a false negative rate of 8%.
- a) Put this in the context of a hypothesis test. What are the null and alternative hypotheses?
  - b) What would a Type I error mean?
  - c) What would a Type II error mean?
  - d) Which is worse here, a Type I or Type II error? Explain.
  - e) What is the power of this test?
- 21. Testing cars.** A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Many states require that cars be tested annually to be sure they meet these standards. Suppose state regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop's license if they find significant evidence that the shop is certifying vehicles that do not meet standards.
- a) In this context, what is a Type I error?
  - b) In this context, what is a Type II error?
  - c) Which type of error would the shop's owner consider more serious?
  - d) Which type of error might environmentalists consider more serious?
- 22. Quality control.** Production managers on an assembly line must monitor the output to be sure that the level of defective products remains small. They periodically inspect a random sample of the items produced. If they find a significant increase in the proportion of items that must be rejected, they will halt the assembly process until the problem can be identified and repaired.
- a) In this context, what is a Type I error?
  - b) In this context, what is a Type II error?
  - c) Which type of error would the factory owner consider more serious?
  - d) Which type of error might customers consider more serious?

23. **Cars again.** As in Exercise 21, state regulators are checking up on repair shops to see if they are certifying vehicles that do not meet pollution standards.
- In this context, what is meant by the power of the test the regulators are conducting?
  - Will the power be greater if they test 20 or 40 cars? Why?
  - Will the power be greater if they use a 5% or a 10% level of significance? Why?
  - Will the power be greater if the repair shop's inspectors are only a little out of compliance or a lot? Why?
24. **Production.** Consider again the task of the quality control inspectors in Exercise 22.
- In this context, what is meant by the power of the test the inspectors conduct?
  - They are currently testing 5 items each hour. Someone has proposed that they test 10 instead. What are the advantages and disadvantages of such a change?
  - Their test currently uses a 5% level of significance. What are the advantages and disadvantages of changing to an alpha level of 1%?
  - Suppose that, as a day passes, one of the machines on the assembly line produces more and more items that are defective. How will this affect the power of the test?
25. **Equal opportunity?** A company is sued for job discrimination because only 19% of the newly hired candidates were minorities when 27% of all applicants were minorities. Is this strong evidence that the company's hiring practices are discriminatory?
- Is this a one-tailed or a two-tailed test? Why?
  - In this context, what would a Type I error be?
  - In this context, what would a Type II error be?
  - In this context, what is meant by the power of the test?
  - If the hypothesis is tested at the 5% level of significance instead of 1%, how will this affect the power of the test?
  - The lawsuit is based on the hiring of 37 employees. Is the power of the test higher than, lower than, or the same as it would be if it were based on 87 hires?
26. **Stop signs.** Highway safety engineers test new road signs, hoping that increased reflectivity will make them more visible to drivers. Volunteers drive through a test course with several of the new- and old-style signs and rate which kind shows up the best.
- Is this a one-tailed or a two-tailed test? Why?
  - In this context, what would a Type I error be?
  - In this context, what would a Type II error be?
  - In this context, what is meant by the power of the test?
  - If the hypothesis is tested at the 1% level of significance instead of 5%, how will this affect the power of the test?
  - The engineers hoped to base their decision on the reactions of 50 drivers, but time and budget constraints may force them to cut back to 20. How would this affect the power of the test? Explain.
27. **Dropouts.** A Statistics professor has observed that for several years about 13% of the students who initially enroll in his Introductory Statistics course withdraw before the end of the semester. A salesman suggests that he try a statistics software package that gets students more involved with computers, predicting that it will cut the dropout rate. The software is expensive, and the salesman offers to let the professor use it for a semester to see if the dropout rate goes down significantly. The professor will have to pay for the software only if he chooses to continue using it.
- Is this a one-tailed or two-tailed test? Explain.
  - Write the null and alternative hypotheses.
  - In this context, explain what would happen if the professor makes a Type I error.
  - In this context, explain what would happen if the professor makes a Type II error.
  - What is meant by the power of this test?
28. **Ads.** A company is willing to renew its advertising contract with a local radio station only if the station can prove that more than 20% of the residents of the city have heard the ad and recognize the company's product. The radio station conducts a random phone survey of 400 people.
- What are the hypotheses?
  - The station plans to conduct this test using a 10% level of significance, but the company wants the significance level lowered to 5%. Why?
  - What is meant by the power of this test?
  - For which level of significance will the power of this test be higher? Why?
  - They finally agree to use  $\alpha = 0.05$ , but the company proposes that the station call 600 people instead of the 400 initially proposed. Will that make the risk of Type II error higher or lower? Explain.
29. **Dropouts, part II.** Initially, 203 students signed up for the Stats course in Exercise 27. They used the software suggested by the salesman, and only 11 dropped out of the course.
- Should the professor spend the money for this software? Support your recommendation with an appropriate test.
  - Explain what your P-value means in this context.
30. **Testing the ads.** The company in Exercise 28 contacts 600 people selected at random, and only 133 remember the ad.
- Should the company renew the contract? Support your recommendation with an appropriate test.
  - Explain what your P-value means in this context.
31. **Two coins.** In a drawer are two coins. They look the same, but one coin produces heads 90% of the time when spun while the other one produces heads only 30% of the time. You select one of the coins. You are allowed to spin it *once* and then must decide whether the coin is the 90%- or the 30%-head coin. Your null hypothesis is that your coin produces 90% heads.
- What is the alternative hypothesis?
  - Given that the outcome of your spin is tails, what would you decide? What if it were heads?
  - How large is  $\alpha$  in this case?
  - How large is the power of this test? (*Hint*: How many possibilities are in the alternative hypothesis?)
  - How could you lower the probability of a Type I error and increase the power of the test at the same time?

32. **Faulty or not?** You are in charge of shipping computers to customers. You learn that a faulty disk drive was put into some of the machines. There's a simple test you can perform, but it's not perfect. All but 4% of the time, a good disk drive passes the test, but unfortunately, 35% of the bad disk drives pass the test, too. You have to decide on the basis of one test whether the disk drive is good or bad. Make this a hypothesis test.
- What are the null and alternative hypotheses?
  - Given that a computer fails the test, what would you decide? What if it passes the test?
  - How large is  $\alpha$  for this test?
  - What is the power of this test? (*Hint:* How many possibilities are in the alternative hypothesis?)
33. **Hoops.** A basketball player with a poor foul-shot record practices intensively during the off-season. He tells the coach that he has raised his proficiency from 60% to 80%. Dubious, the coach asks him to take 10 shots, and is surprised when the player hits 9 out of 10. Did the player prove that he has improved?
- Suppose the player really is no better than before—still a 60% shooter. What's the probability he can hit at least 9 of 10 shots anyway? (*Hint:* Use a Binomial model.)
  - If that is what happened, now the coach thinks the player has improved when he has not. Which type of error is that?
  - If the player really can hit 80% now, and it takes at least 9 out of 10 successful shots to convince the coach, what's the power of the test?
  - List two ways the coach and player could increase the power to detect any improvement.
34. **Pottery.** An artist experimenting with clay to create pottery with a special texture has been experiencing difficulty with these special pieces. About 40% break in the kiln during firing. Hoping to solve this problem, she buys some more expensive clay from another supplier. She plans to make and fire 10 pieces and will decide to use the new clay if at most one of them breaks.
- Suppose the new, expensive clay really is no better than her usual clay. What's the probability that this test convinces her to use it anyway? (*Hint:* Use a Binomial model.)
  - If she decides to switch to the new clay and it is no better, what kind of error did she commit?
  - If the new clay really can reduce breakage to only 20%, what's the probability that her test will not detect the improvement?
  - How can she improve the power of her test? Offer at least two suggestions.



### JUST CHECKING Answers

- With a z-score of 0.62, you can't reject the null hypothesis. The experiment shows no evidence that the wheel is not fair.
- At  $\alpha = 0.05$ , you can't reject the null hypothesis because 0.30 is contained in the 90% confidence interval—it's plausible that sending the DVDs is no more effective than just sending letters.
- The confidence interval is from 29% to 45%. The DVD strategy is more expensive and may not be worth it. We can't distinguish the success rate from 30% given the results of this experiment, but 45% would represent a large improvement. The bank should consider another trial, increasing their sample size to get a narrower confidence interval.
- A Type I error would mean deciding that the DVD success rate is higher than 30% when it really isn't. They would adopt a more expensive method for collecting payments that's no better than the less expensive strategy.
- A Type II error would mean deciding that there's not enough evidence to say that the DVD strategy works when in fact it does. The bank would fail to discover an effective method for increasing their revenue from delinquent accounts.
- 60%; the larger the effect size, the greater the power. It's easier to detect an improvement to a 60% success rate than to a 32% rate.

# Comparing Two Proportions



**WHO** 6971 male drivers  
**WHAT** Seatbelt use  
**WHY** Highway safety  
**WHEN** 2007  
**WHERE** Massachusetts

**D**o men take more risks than women? Psychologists have documented that in many situations, men choose riskier behavior than women do. But what is the effect of having a woman by their side? A recent seatbelt observation study in Massachusetts<sup>1</sup> found that, not surprisingly, male drivers wear seatbelts less often than women do. The study also noted that men's belt-wearing jumped more than 16 percentage points when they had a female passenger. Seatbelt use was recorded at 161 locations in Massachusetts, using random-sampling methods developed by the National Highway Traffic Safety Administration (NHTSA). Female drivers wore belts more than 70% of the time, regardless of the sex of their passengers. Of 4208 male drivers with female passengers, 2777 (66.0%) were belted. But among 2763 male drivers with male passengers only, 1363 (49.3%) wore seatbelts. This was only a random sample, but it suggests there may be a shift in men's risk-taking behavior when women are present. What would we estimate the true size of that gap to be?

Comparisons between two percentages are much more common than questions about isolated percentages. And they are more interesting. We often want to know how two groups differ, whether a treatment is better than a placebo control, or whether this year's results are better than last year's.

## Another Ruler

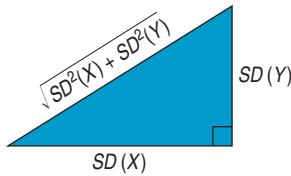
We know the difference between the proportions of men wearing seatbelts seen in the *sample*. It's 16.7%. But what's the *true* difference for all men? We know that our estimate probably isn't exactly right. To say more, we need a new ruler—the standard deviation of the sampling distribution model for the difference in the proportions. Now we have two proportions, and each will vary from sample to sample. We are interested in the difference between them. So what is the correct standard deviation?

<sup>1</sup>Massachusetts Traffic Safety Research Program [June 2007].



The answer comes to us from Chapter 16. Remember the Pythagorean Theorem of Statistics?

*The variance of the sum or difference of two independent random variables is the sum of their variances.*



This is such an important (and powerful) idea in Statistics that it's worth pausing a moment to review the reasoning. Here's some intuition about why variation increases even when we subtract two random quantities.

Grab a full box of cereal. The box claims to contain 16 ounces of cereal. We know that's not exact: There's some small variation from box to box. Now pour a bowl of cereal. Of course, your 2-ounce serving will not be exactly 2 ounces. There'll be some variation there, too. How much cereal would you guess was left in the box? Do you think your guess will be as close as your guess for the full box? After you pour your bowl, the amount of cereal in the box is still a random quantity (with a smaller mean than before), but it is even *more variable* because of the additional variation in the amount you poured.

According to our rule, the variance of the amount of cereal left in the box would now be the *sum* of the two *variances*.

We want a standard deviation, not a variance, but that's just a square root away. We can write symbolically what we've just said:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y), \text{ so}$$

$$SD(X - Y) = \sqrt{SD^2(X) + SD^2(Y)} = \sqrt{\text{Var}(X) + \text{Var}(Y)}.$$

Be careful, though—this simple formula applies only when  $X$  and  $Y$  are independent. Just as the Pythagorean Theorem<sup>2</sup> works only for right triangles, our formula works only for independent random variables. Always check for independence before using it.

For independent random variables, variances add.

## The Standard Deviation of the Difference Between Two Proportions

Combining independent random quantities always *increases* the overall variation, so even for *differences* of independent random variables, variances add.

Fortunately, proportions observed in independent random samples *are* independent, so we can put the two proportions in for  $X$  and  $Y$  and add their variances. We just need to use careful notation to keep things straight.

When we have two samples, each can have a different size and proportion value, so we keep them straight with subscripts. Often we choose subscripts that remind us of the groups. For our example, we might use "M" and "F", but generically we'll just use "1" and "2". We will represent the two sample proportions as  $\hat{p}_1$  and  $\hat{p}_2$ , and the two sample sizes as  $n_1$  and  $n_2$ .

The standard deviations of the sample proportions are  $SD(\hat{p}_1) = \sqrt{\frac{p_1q_1}{n_1}}$  and

$SD(\hat{p}_2) = \sqrt{\frac{p_2q_2}{n_2}}$ , so the variance of the difference in the proportions is

$$\text{Var}(\hat{p}_1 - \hat{p}_2) = \left(\sqrt{\frac{p_1q_1}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2q_2}{n_2}}\right)^2 = \frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}.$$

The standard deviation is the square root of that variance:

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}.$$



<sup>2</sup>If you don't remember the formula, don't rely on the Scarecrow's version from *The Wizard of Oz*. He may have a brain and have been awarded his Th.D. (Doctor of Thinkology), but he gets the formula wrong.

We usually don't know the true values of  $p_1$  and  $p_2$ . When we have the sample proportions in hand from the data, we use them to estimate the variances. So the standard error is

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

### FOR EXAMPLE

#### Finding the standard error of a difference in proportions

A recent survey of 886 randomly selected teenagers (aged 12–17) found that more than half of them had online profiles.<sup>3</sup> Some researchers and privacy advocates are concerned about the possible access to personal information about teens in public places on the Internet. There appear to be differences between boys and girls in their online behavior. Among teens aged 15–17, 57% of the 248 boys had posted profiles, compared to 70% of the 256 girls. Let's start the process of estimating how large the true gender gap might be.

**Question:** What's the standard error of the difference in sample proportions?

Because the boys and girls were selected at random, it's reasonable to assume their behaviors are independent, so it's okay to use the Pythagorean Theorem of Statistics and add the variances:

$$\begin{aligned} SE(\hat{p}_{\text{boys}}) &= \sqrt{\frac{0.57 \times 0.43}{248}} = 0.0314 & SE(\hat{p}_{\text{girls}}) &= \sqrt{\frac{0.70 \times 0.30}{256}} = 0.0286 \\ SE(\hat{p}_{\text{girls}} - \hat{p}_{\text{boys}}) &= \sqrt{0.0314^2 + 0.0286^2} = 0.0425 \end{aligned}$$

## Assumptions and Conditions

Before we look at our example, we need to check assumptions and conditions.

### INDEPENDENCE ASSUMPTIONS

**Independence Assumption:** Within each group, the data should be based on results for independent individuals. We can't check that for certain, but we *can* check the following:

**Randomization Condition:** The data in each group should be drawn independently and at random from a homogeneous population or generated by a randomized comparative experiment.

**The 10% Condition:** If the data are sampled without replacement, the sample should not exceed 10% of the population.

Because we are comparing two groups in this way, we need an additional Independence Assumption. In fact, this is the most important of these assumptions. If it is violated, these methods just won't work.

**Independent Groups Assumption:** The two groups we're comparing must also be independent of *each other*. Usually, the independence of the groups from each other is evident from the way the data were collected.

Why is the Independent Groups Assumption so important? If we compare husbands with their wives, or a group of subjects before and after some treatment, we can't just add the variances. Subjects' performance before a treatment might very well be related to their performance after the treatment. So the proportions are not independent and the Pythagorean-style variance formula does not hold. We'll see a way to compare a common kind of nonindependent samples in a later chapter.

<sup>3</sup> Princeton Survey Research Associates International for the Pew Internet & American Life Project.

## SAMPLE SIZE CONDITION

Each of the groups must be big enough. As with individual proportions, we need larger groups to estimate proportions that are near 0% or 100%. We usually check the Success/Failure Condition for each group.

**Success/Failure Condition:** Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each.

### FOR EXAMPLE

#### Checking assumptions and conditions

**Recap:** Among randomly sampled teens aged 15–17, 57% of the 248 boys had posted online profiles, compared to 70% of the 256 girls.

**Question:** Can we use these results to make inferences about all 15–17-year-olds?

- ✓ **Randomization Condition:** The sample of boys and the sample of girls were both chosen randomly.
- ✓ **10% Condition:** 248 boys and 256 girls are each less than 10% of all teenage boys and girls.
- ✓ **Independent Groups Assumption:** Because the samples were selected at random, it's reasonable to believe the boys' online behaviors are independent of the girls' online behaviors.
- ✓ **Success/Failure Condition:** Among the boys,  $248(0.57) = 141$  had online profiles and the other  $248(0.43) = 107$  did not. For the girls,  $256(0.70) = 179$  successes and  $256(0.30) = 77$  failures. All counts are at least 10.

Because all the assumptions and conditions are satisfied, it's okay to proceed with inference for the difference in proportions.

(Note that when we find the *observed* counts of successes and failures, we round off to whole numbers. We're using the reported percentages to recover the actual counts.)

## The Sampling Distribution

We're almost there. We just need one more fact about proportions. We already know that for large enough samples, each of our proportions has an approximately Normal sampling distribution. The same is true of their difference.

### Why Normal?

In Chapter 16 we learned that sums and differences of independent Normal random variables also follow a Normal model. That's the reason we use a Normal model for the difference of two independent proportions.

### THE SAMPLING DISTRIBUTION MODEL FOR A DIFFERENCE BETWEEN TWO INDEPENDENT PROPORTIONS

Provided that the sampled values are independent, the samples are independent, and the sample sizes are large enough, the sampling distribution of  $\hat{p}_1 - \hat{p}_2$  is modeled by a Normal model with mean  $\mu = p_1 - p_2$  and standard deviation

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

The sampling distribution model and the standard deviation give us all we need to find a margin of error for the difference in proportions—or at least they would if we knew the true proportions,  $p_1$  and  $p_2$ . However, we don't know the true values, so we'll work with the observed proportions,  $\hat{p}_1$  and  $\hat{p}_2$ , and use  $SE(\hat{p}_1 - \hat{p}_2)$  to estimate the standard deviation. The rest is just like a one-proportion z-interval.

**AS** **Activity: Compare Two Proportions.** Does a preschool program help disadvantaged children later in life?

### A TWO-PROPORTION z-INTERVAL

When the conditions are met, we are ready to find the confidence interval for the difference of two proportions,  $p_1 - p_2$ . The confidence interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

where we find the standard error of the difference,

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

from the observed proportions.

The critical value  $z^*$  depends on the particular confidence level,  $C$ , that we specify.

### FOR EXAMPLE

#### Finding a two-proportion z-interval

**Recap:** Among randomly sampled teens aged 15–17, 57% of the 248 boys had posted online profiles, compared to 70% of the 256 girls. We calculated the standard error for the difference in sample proportions to be  $SE(\hat{p}_{girls} - \hat{p}_{boys}) = 0.0425$  and found that the assumptions and conditions required for inference checked out okay.

**Question:** What does a confidence interval say about the difference in online behavior?

A 95% confidence interval for  $p_{girls} - p_{boys}$  is  $(\hat{p}_{girls} - \hat{p}_{boys}) \pm z^*SE(\hat{p}_{girls} - \hat{p}_{boys})$

$$(0.70 - 0.57) \pm 1.96(0.0425)$$

$$0.13 \pm 0.083$$

$$(4.7\%, 21.3\%)$$

We can be 95% confident that among teens aged 15–17, the proportion of girls who post online profiles is between 4.7 and 21.3 percentage points higher than the proportion of boys who do. It seems clear that teen girls are more likely to post profiles than are boys the same age.

### STEP-BY-STEP EXAMPLE

#### A Two-Proportion z-Interval

Now we are ready to be more precise about the passenger-based gap in male drivers' seatbelt use. We'll estimate the difference with a confidence interval using a method called the **two-proportion z-interval** and follow the four confidence interval steps.

**Question:** How much difference is there in the proportion of male drivers who wear seatbelts when sitting next to a male passenger and the proportion who wear seatbelts when sitting next to a female passenger?



**Plan** State what you want to know. Discuss the variables and the W's.

Identify the parameter you wish to estimate. (It usually doesn't matter in which direction we subtract, so, for convenience, we usually choose the direction with a positive difference.)

I want to know the true difference in the population proportion,  $p_M$ , of male drivers who wear seatbelts when sitting next to a man and  $p_F$ , the proportion who wear seatbelts when sitting next to a woman. The data are from a random sample of drivers in Massachusetts in 2007, observed according to procedures developed by the NHTSA. The parameter of interest is the difference  $p_F - p_M$ .

Choose and state a confidence level.

**Model** Think about the assumptions and check the conditions.

The Success/Failure Condition must hold for each group.

State the sampling distribution model for the statistic.

Choose your method.

I will find a 95% confidence interval for this parameter.

- ✓ **Independence Assumption:** Driver behavior was independent from car to car.
- ✓ **Randomization Condition:** The NHTSA methods are more complex than an SRS, but they result in a suitable random sample.
- ✓ **10% Condition:** The samples include far fewer than 10% of all male drivers accompanied by male or by female passengers.
- ✓ **Independent Groups Assumption:** There's no reason to believe that seatbelt use among drivers with male passengers and those with female passengers are not independent.
- ✓ **Success Failure Condition:** Among male drivers with female passengers, 2777 wore seatbelts and 1431 did not; of those driving with male passengers, 1363 wore seatbelts and 1400 did not. Each group contained far more than 10 successes and 10 failures.

Under these conditions, the sampling distribution of the difference between the sample proportions is approximately Normal, so I'll find a **two-proportion z-interval**.



**Mechanics** Construct the confidence interval.

As often happens, the key step in finding the confidence interval is estimating the standard deviation of the sampling distribution model of the statistic. Here the statistic is the difference in the proportions of men who wear seatbelts when they have a female passenger and the proportion who do so with a male passenger. Substitute the data values into the formula.

The sampling distribution is Normal, so the critical value for a 95% confidence interval,  $z^*$ , is 1.96. The margin of error is the critical value times the SE.

I know

$$n_F = 4208, n_M = 2763.$$

The observed sample proportions are

$$\hat{p}_F = \frac{2777}{4208} = 0.660, \hat{p}_M = \frac{1363}{2763} = 0.493$$

I'll estimate the SD of the difference with

$$\begin{aligned} SE(\hat{p}_F - \hat{p}_M) &= \sqrt{\frac{\hat{p}_F \hat{q}_F}{n_F} + \frac{\hat{p}_M \hat{q}_M}{n_M}} \\ &= \sqrt{\frac{(0.660)(0.340)}{4208} + \frac{(0.493)(0.507)}{2763}} \\ &= 0.012 \end{aligned}$$

$$\begin{aligned} ME &= z^* \times SE(\hat{p}_F - \hat{p}_M) \\ &= 1.96(0.012) = 0.024 \end{aligned}$$

The confidence interval is the statistic  $\pm$ ME.

The observed difference in proportions is  $\hat{p}_F - \hat{p}_M = 0.660 - 0.493 = 0.167$ , so the 95% confidence interval is

$$0.167 \pm 0.024$$

or 14.3% to 19.1%



**Conclusion** Interpret your confidence interval in the proper context. (Remember: We're 95% confident that our interval captured the true difference.)

I am 95% confident that the proportion of male drivers who wear seatbelts when driving next to a female passenger is between 14.3 and 19.1 percentage points higher than the proportion who wear seatbelts when driving next to a male passenger.

This is an interesting result—but be careful not to try to say too much! In Massachusetts, overall seatbelt use is lower than the national average, so we can't be certain that these results generalize to other states. And these were two different groups of men, so we can't say that, individually, men are more likely to buckle up when they have a woman passenger. You can probably think of several alternative explanations; we'll suggest just a couple. Perhaps age is a lurking variable: Maybe older men are more likely to wear seatbelts and also more likely to be driving with their wives. Or maybe men who don't wear seatbelts have trouble attracting women!

### TI Tips

### Finding a confidence interval

You can use a routine in the **STAT TESTS** menu to create confidence intervals for the difference of two proportions. Remember, the calculator can do only the mechanics—checking conditions and writing conclusions are still up to you.

A Gallup Poll asked whether the attribute “intelligent” described men in general. The poll revealed that 28% of 506 men thought it did, but only 14% of 520 women agreed. We want to estimate the true size of the gender gap by creating a 95% confidence interval.

- Go to the **STAT TESTS** menu. Scroll down the list and select **B: 2-PropZInt**.
- Enter the observed number of males: **.28\*506**. Remember that the actual number of males must be a whole number, so be sure to round off.
- Enter the sample size: **506** males.
- Repeat those entries for women: **.14\*520** agreed, and the sample size was **520**.
- Specify the desired confidence level.
- **Calculate** the result.

And now explain what you see: We are 95% confident that the proportion of men who think the attribute “intelligent” describe males in general is between 9 and 19 percentage points higher than the proportion of women who think so.

```
EDIT CALC TESTS
012-SampTInt...
A:1-PropZInt...
B:2-PropZInt...
C:x2-Test...
D:x2GOF-Test...
E:2-SampFTest...
F↓LinRegTTest...
```

```
2-PropZInt
x1:142
n1:506
x2:73
n2:520
C-Level: .95
Calculate
```

```
2-PropZInt
(.09101, .18948)
p1=.2806324111
p2=.1403846154
n1=506
n2=520
```



## JUST CHECKING

A public broadcasting station plans to launch a special appeal for additional contributions from current members. Unsure of the most effective way to contact people, they run an experiment. They randomly select two groups of current members. They send the same request for donations to everyone, but it goes to one group by e-mail and to the other group by regular mail. The station was successful in getting contributions from 26% of the members they e-mailed but only from 15% of those who received the request by regular mail. A 90% confidence interval estimated the difference in donation rates to be  $11\% \pm 7\%$ .

1. Interpret the confidence interval in this context.
2. Based on this confidence interval, what conclusion would we reach if we tested the hypothesis that there's no difference in the response rates to the two methods of fundraising? Explain.

## Will I Snore When I'm 64?

<b>WHO</b>	Randomly selected U.S. adults over age 18
<b>WHAT</b>	Proportion who snore, categorized by age (less than 30, 30 or older)
<b>WHEN</b>	2001
<b>WHERE</b>	United States
<b>WHY</b>	To study sleep behaviors of U.S. adults



The National Sleep Foundation asked a random sample of 1010 U.S. adults questions about their sleep habits. The sample was selected in the fall of 2001 from random telephone numbers, stratified by region and sex, guaranteeing that an equal number of men and women were interviewed (2002 Sleep in America Poll, National Sleep Foundation, Washington, DC).

One of the questions asked about snoring. Of the 995 respondents, 37% of adults reported that they snored at least a few nights a week during the past year. Would you expect that percentage to be the same for all age groups? Split into two age categories, 26% of the 184 people under 30 snored, compared with 39% of the 811 in the older group. Is this difference of 13% real, or due only to natural fluctuations in the sample we've chosen?

The question calls for a hypothesis test. Now the parameter of interest is the true *difference* between the (reported) snoring rates of the two age groups.

What's the appropriate null hypothesis? That's easy here. We hypothesize that there is no difference in the proportions. This is such a natural null hypothesis that we rarely consider any other. But instead of writing  $H_0: p_1 = p_2$ , we usually express it in a slightly different way. To make it relate directly to the *difference*, we hypothesize that the difference in proportions is zero:

$$H_0: p_1 - p_2 = 0.$$

## Everyone into the Pool

Our hypothesis is about a new parameter: the *difference* in proportions. We'll need a standard error for that. Wait—don't we know that already? Yes and no. We know that the standard error of the difference in proportions is

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}},$$

and we could just plug in the numbers, but we can do even better. The secret is that proportions and their standard deviations are linked. There are two proportions in the standard error formula—but look at the null hypothesis. It says that these proportions are equal. To do a hypothesis test, we *assume* that the null hypothesis is true. So there should be just a single value of  $\hat{p}$  in the SE formula (and, of course,  $\hat{q}$  is just  $1 - \hat{p}$ ).

How would we do this for the snoring example? If the null hypothesis is true, then, among all adults, the two groups have the same proportion. Overall, we saw  $48 + 318 = 366$  snorers out of a total of  $184 + 811 = 995$  adults who responded to this question. The overall proportion of snorers was  $366/995 = 0.3678$ .

Combining the counts like this to get an overall proportion is called **pooling**. Whenever we have data from different sources or different groups but we believe that they really came from the same underlying population, we pool them to get better estimates.

When we have counts for each group, we can find the pooled proportion as

$$\hat{p}_{\text{pooled}} = \frac{\text{Success}_1 + \text{Success}_2}{n_1 + n_2},$$

where  $\text{Success}_1$  is the number of successes in group 1 and  $\text{Success}_2$  is the number of successes in group 2. That's the overall proportion of success.

When we have only proportions and not the counts, as in the snoring example, we have to reconstruct the number of successes by multiplying the sample sizes by the proportions:

$$\text{Success}_1 = n_1\hat{p}_1 \quad \text{and} \quad \text{Success}_2 = n_2\hat{p}_2.$$

If these calculations don't come out to whole numbers, round them first. There must have been a whole number of successes, after all. (This is the *only* time you should round values in the middle of a calculation.)

We then put this pooled value into the formula, substituting it for *both* sample proportions in the standard error formula:

$$\begin{aligned} SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) &= \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}} \\ &= \sqrt{\frac{0.3678 \times (1 - 0.3678)}{184} + \frac{0.3678 \times (1 - 0.3678)}{811}}. \end{aligned}$$

This comes out to 0.039.

When finding the number of successes, round the values to integers. For example, the 48 snorers among the 184 under-30 respondents are actually 26.1% of 184. We round back to the nearest whole number to find the count that could have yielded the rounded percent we were given.

## Improving the Success/Failure Condition

The vaccine Gardasil<sup>®</sup> was introduced to prevent the strains of human papillomavirus (HPV) that are responsible for almost all cases of cervical cancer. In randomized placebo-controlled clinical trials,<sup>4</sup> only 1 case of HPV was diagnosed among 7897 women who received the vaccine, compared with 91 cases diagnosed among 7899 who received a placebo. The one observed HPV case ("success") doesn't meet the at-least-10-successes criterion. Surely, though, we should not refuse to test the effectiveness of the vaccine just because it failed so rarely; that would be absurd.

For that reason, in a two-proportion  $z$ -test, the proper Success/Failure test uses the *expected* frequencies, which we can find from the pooled proportion. In this case,

$$\begin{aligned} \hat{p}_{\text{pooled}} &= \frac{91 + 1}{7899 + 7897} = 0.0058 \\ n_1\hat{p}_{\text{pooled}} &= 7899(0.0058) = 46 \\ n_2\hat{p}_{\text{pooled}} &= 7897(0.0058) = 46, \end{aligned}$$

so we can proceed with the hypothesis test.

<sup>4</sup> *Quadrivalent Human Papillomavirus Vaccine: Recommendations of the Advisory Committee on Immunization Practices (ACIP)*, National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention [May 2007].



Often it is easier just to check the observed numbers of successes and failures. If they are both greater than 10, you don't need to look further. But keep in mind that the correct test uses the expected frequencies rather than the observed ones.

## Compared to What?

Naturally, we'll reject our null hypothesis if we see a large enough difference in the two proportions. How can we decide whether the difference we see,  $\hat{p}_1 - \hat{p}_2$ , is large? The answer is the same as always: We just compare it to its standard deviation.

Unlike previous hypothesis-testing situations, the null hypothesis doesn't provide a standard deviation, so we'll use a standard error (here, pooled). Since the sampling distribution is Normal, we can divide the observed difference by its standard error to get a z-score. The z-score will tell us how many standard errors the observed difference is away from 0. We can then use the 68–95–99.7 Rule to decide whether this is large, or some technology to get an exact P-value. The result is a **two-proportion z-test**.

**AS** **Activity: Test for a Difference Between Two Proportions.** Is premium-brand chicken less likely to be contaminated than store-brand chicken?

### TWO-PROPORTION z-TEST

The conditions for the two-proportion z-test are the same as for the two-proportion z-interval. We are testing the hypothesis

$$H_0: p_1 - p_2 = 0.$$

Because we hypothesize that the proportions are equal, we pool the groups to find

$$\hat{p}_{\text{pooled}} = \frac{\text{Success}_1 + \text{Success}_2}{n_1 + n_2}$$

and use that pooled value to estimate the standard error:

$$SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}}.$$

Now we find the test statistic,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)}.$$

When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

### STEP-BY-STEP EXAMPLE

### A Two-Proportion z-Test

**Question:** Are the snoring rates of the two age groups really different?



**Plan** State what you want to know. Discuss the variables and the W's.

I want to know whether snoring rates differ for those under and over 30 years old. The data are from a random sample of 1010 U.S. adults surveyed in the 2002 Sleep in America Poll. Of these, 995 responded to the question about snoring, indicating whether or not they had snored at least a few nights a week in the past year.

**Hypotheses** The study simply broke down the responses by age, so there is no sense that either alternative was preferred. A two-sided alternative hypothesis is appropriate.

**Model** Think about the assumptions and check the conditions.

State the null model.

Choose your method.

$H_0$ : There is no difference in snoring rates in the two age groups:

$$p_{old} - p_{young} = 0.$$

$H_A$ : The rates are different:  $p_{old} - p_{young} \neq 0$ .

- ✓ **Independence Assumption:** The National Sleep Foundation selected respondents at random, so they should be independent.
- ✓ **Randomization Condition:** The respondents were randomly selected by telephone number and stratified by sex and region.
- ✓ **10% Condition:** The number of adults surveyed in each age group is certainly far less than 10% of that population.
- ✓ **Independent Groups Assumption:** The two groups are independent of each other because the sample was selected at random.
- ✓ **Success/Failure Condition:** In the younger age group, 48 snored and 136 didn't. In the older group, 318 snored and 493 didn't. The observed numbers of both successes and failures are much more than 10 for both groups.<sup>5</sup>

Because the conditions are satisfied, I'll use a Normal model and perform a **two-proportion z-test**.



### Mechanics

The hypothesis is that the proportions are equal, so pool the sample data.

Use the pooled SE to estimate  $SD(p_{old} - p_{young})$ .

$$n_{young} = 184, y_{young} = 48, \hat{p}_{young} = 0.261$$

$$n_{old} = 811, y_{old} = 318, \hat{p}_{old} = 0.392$$

$$\hat{p}_{pooled} = \frac{y_{old} + y_{young}}{n_{old} + n_{young}} = \frac{318 + 48}{811 + 184} = 0.3678$$

$$SE_{pooled}(\hat{p}_{old} - \hat{p}_{young})$$

$$= \sqrt{\frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_{old}} + \frac{\hat{p}_{pooled}\hat{q}_{pooled}}{n_{young}}}$$

$$= \sqrt{\frac{(0.3678)(0.6322)}{811} + \frac{(0.3678)(0.6322)}{184}}$$

$$\approx 0.039375$$

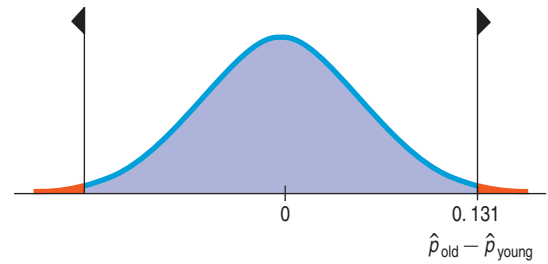
The observed difference in sample proportions is  $\hat{p}_{old} - \hat{p}_{young} = 0.392 - 0.261 = 0.131$

<sup>5</sup>This is one of those situations in which the traditional term “success” seems a bit weird. A success here could be that a person snores. “Success” and “failure” are arbitrary labels left over from studies of gambling games.

Make a picture. Sketch a Normal model centered at the hypothesized difference of 0. Shade the region to the right of the observed difference, and because this is a two-tailed test, also shade the corresponding region in the other tail.

Find the z-score for the observed difference in proportions, 0.131.

Find the P-value using Table Z or technology. Because this is a two-tailed test, we must *double* the probability we find in the upper tail.



$$z = \frac{(\hat{p}_{old} - \hat{p}_{young}) - 0}{SE_{pooled}(\hat{p}_{old} - \hat{p}_{young})} = \frac{0.131 - 0}{0.039375} = 3.33$$

$$P = 2P(z \geq 3.33) = 0.0008$$



**Conclusion** Link the P-value to your decision about the null hypothesis, and state your conclusion in context.

The P-value of 0.0008 says that if there really were no difference in (reported) snoring rates between the two age groups, then the difference observed in this study would happen only 8 times in 10,000. This is so small that I reject the null hypothesis of no difference and conclude that there is a difference in the rate of snoring between older adults and younger adults. It appears that older adults are more likely to snore.

## TI Tips

## Testing the hypothesis

```
EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
```

```
2-PropZTest
x1:318
n1:811
x2:48
n2:184
P1: <P2 >P2
Calculate Draw
```

```
2-PropZTest
P1≠P2
z=3.332941852
P=8.5944146E-4
p1=.392103508
p2=.2688895652
↓P=.367839196
```

Yes, of course, there's a **STAT TESTS** routine to test a hypothesis about the difference of two proportions. Let's do the mechanics for the test about snoring. Of 811 people over 30 years old, 318 snored, while only 48 of the 184 people under 30 did.

- In the **STAT TESTS** menu select **6:2-PropZTest**.
- Enter the observed numbers of snorers and the sample sizes for both groups.
- Since this is a two-tailed test, indicate that you want to see if the proportions are unequal. When you choose this option, the calculator will automatically include both tails as it determines the P-value.
- **Calculate** the result. Don't worry; for this procedure the calculator will pool the proportions automatically.

Now it is up to you to interpret the result and state a conclusion. We see a z-score of 3.33 and the P-value is 0.0008. Such a small P-value indicates that the observed difference is unlikely to be sampling error. What does that mean about snoring and age? Here's a great opportunity to follow up with a confidence interval so you can Tell even more!



## JUST CHECKING

- A June 2004 public opinion poll asked 1000 randomly selected adults whether the United States should decrease the amount of immigration allowed; 49% of those responding said “yes.” In June of 1995, a random sample of 1000 had found that 65% of adults thought immigration should be curtailed. To see if that percentage has decreased, why can’t we just use a one-proportion z-test of  $H_0: p = 0.65$  and see what the P-value for  $\hat{p} = 0.49$  is?
- For opinion polls like this, which has more variability: the percentage of respondents answering “yes” in either year or the difference in the percentages between the two years?

## FOR EXAMPLE

### Another 2-proportion z-test

**Recap:** One concern of the study on teens’ online profiles was safety and privacy. In the random sample, girls were less likely than boys to say that they are easy to find online from their profiles. Only 19% (62 girls) of 325 teen girls with profiles say that they are easy to find, while 28% (75 boys) of the 268 boys with profiles say the same.

**Question:** Are these results evidence of a real difference between boys and girls? Perform a two-proportion z-test and discuss what you find.

$$H_0: p_{\text{boys}} - p_{\text{girls}} = 0$$

$$H_A: p_{\text{boys}} - p_{\text{girls}} \neq 0$$

- ✓ **Randomization Condition:** The sample of boys and the sample of girls were both chosen randomly.
- ✓ **10% Condition:** 268 boys and 325 girls are each less than 10% of all teenage boys and girls with online profiles.
- ✓ **Independent Groups Assumption:** Because the samples were selected at random, it’s reasonable to believe the boys’ perceptions are independent of the girls’.
- ✓ **Success/Failure Condition:** Among the girls, there were 62 “successes” and 263 failures, and among boys, 75 successes and 193 failures. These counts are at least 10 for each group.

Because all the assumptions and conditions are satisfied, it’s okay to do a **two-proportion z-test**:

$$\begin{aligned}\hat{p}_{\text{pooled}} &= \frac{75 + 62}{268 + 325} = 0.231 \\ SE_{\text{pooled}}(\hat{p}_{\text{boys}} - \hat{p}_{\text{girls}}) &= \sqrt{\frac{0.231 \times 0.769}{268} + \frac{0.231 \times 0.769}{325}} = 0.0348 \\ z &= \frac{(0.28 - 0.19) - 0}{0.0348} = 2.59 \\ P(z > 2.59) &= 0.0048\end{aligned}$$


This is a two-tailed test, so the P-value =  $2(0.0048) = 0.0096$ . Because this P-value is very small, I reject the null hypothesis. This study provides strong evidence that there really is a difference in the proportions of teen girls and boys who say they are easy to find online.

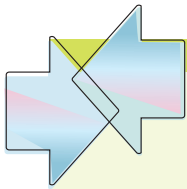
## WHAT CAN GO WRONG?

- ▶ **Don’t use two-sample proportion methods when the samples aren’t independent.** These methods give wrong answers when this assumption of independence is violated. Good random sampling is usually the best insurance of independent groups. Make sure there is no relationship between the two groups. For example, you can’t compare the

proportion of respondents who own SUVs with the proportion of those same respondents who think the tax on gas should be eliminated. The responses are not independent because you've asked the same people. To use these methods to estimate or test the difference, you'd need to survey two different groups of people.

Alternatively, if you have a random sample, you can split your respondents according to their answers to one question and treat the two resulting groups as independent samples. So, you could test whether the proportion of SUV owners who favored eliminating the gas tax was the same as the corresponding proportion among non-SUV owners.

- ▶ **Don't apply inference methods where there was no randomization.** If the data do not come from representative random samples or from a properly randomized experiment, then the inference about the differences in proportions will be wrong.
- ▶ **Don't interpret a significant difference in proportions causally.** It turns out that people with higher incomes are more likely to snore. Does that mean money affects sleep patterns? Probably not. We have seen that older people are more likely to snore, and they are also likely to earn more. In a prospective or retrospective study, there is always the danger that other lurking variables not accounted for are the real reason for an observed difference. Be careful not to jump to conclusions about causality. 



## CONNECTIONS

In Chapter 3 we looked at contingency tables for two categorical variables. Differences in proportions are just  $2 \times 2$  contingency tables. You'll often see data presented in this way. For example, the snoring data could be shown as

	18–29	30 and over	Total
Snore	48	318	366
Don't snore	136	493	629
Total	184	811	995

We tested whether the column percentages of snorers were the same for the two age groups.

This chapter gives the first examples we've seen of inference methods for a parameter other than a simple proportion. Although we have a different standard error, the step-by-step procedures are almost identical. In particular, once again we divide the statistic (the difference in proportions) by its standard error and get a  $z$ -score. You should feel right at home.



## WHAT HAVE WE LEARNED?

In the last few chapters we began our exploration of statistical inference; we learned how to create confidence intervals and test hypotheses about a proportion. Now we've looked at inference for the difference in two proportions. In doing so, perhaps the most important thing we've learned is that the concepts and interpretations are essentially the same—only the mechanics have changed slightly.

We've learned that hypothesis tests and confidence intervals for the difference in two proportions are based on Normal models. Both require us to find the standard error of the difference in

two proportions. We do that by adding the variances of the two sample proportions, assuming our two groups are independent. When we test a hypothesis that the two proportions are equal, we pool the sample data; for confidence intervals, we don't pool.

## Terms

**Variances of independent random variables add**

506. The variance of a sum or difference of independent random variables is the sum of the variances of those variables.

**Sampling distribution of the difference between two proportions**

507. The sampling distribution of  $\hat{p}_1 - \hat{p}_2$  is, under appropriate assumptions, modeled by a Normal model with mean  $\mu = p_1 - p_2$  and standard deviation  $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$ .

**Two-proportion z-interval**

508. A two-proportion z-interval gives a confidence interval for the true difference in proportions,  $p_1 - p_2$ , in two independent groups.

The confidence interval is  $(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$ , where  $z^*$  is a critical value from the standard Normal model corresponding to the specified confidence level.

**Pooling**

512. When we have data from different sources that we believe are homogeneous, we can get a better estimate of the common proportion and its standard deviation. We can combine, or pool, the data into a single group for the purpose of estimating the common proportion. The resulting pooled standard error is based on more data and is thus more reliable (if the null hypothesis is true and the groups are truly homogeneous).

**Two-proportion z-test**

513. Test the null hypothesis  $H_0: p_1 - p_2 = 0$  by referring the statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)}$$

to a standard Normal model.

## Skills

**THINK**

- ▶ Be able to state the null and alternative hypotheses for testing the difference between two population proportions.
- ▶ Know how to examine your data for violations of conditions that would make inference about the difference between two population proportions unwise or invalid.
- ▶ Understand that the formula for the standard error of the difference between two independent sample proportions is based on the principle that when finding the sum or difference of two independent random variables, their variances add.

**SHOW**

- ▶ Know how to find a confidence interval for the difference between two proportions.
- ▶ Be able to perform a significance test of the natural null hypothesis that two population proportions are equal.

**TELL**

- ▶ Know how to write a sentence describing what is said about the difference between two population proportions by a confidence interval.
- ▶ Know how to write a sentence interpreting the results of a significance test of the null hypothesis that two population proportions are equal.
- ▶ Be able to interpret the meaning of a P-value in nontechnical language, making clear that the probability claim is made about computed values and not about the population parameter of interest.
- ▶ Know that we do not “accept” a null hypothesis if we fail to reject it.

## INFERENCES FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS ON THE COMPUTER

It is so common to test against the null hypothesis of no difference between the two true proportions that most statistics programs simply assume this null hypothesis. And most will automatically use the pooled standard deviation. If you wish to test a different null (say, that the true difference is 0.3), you may have to search for a way to do it.

Many statistics packages don't offer special commands for inference for differences between proportions. As with inference for single proportions, most statistics programs want the "success" and "failure" status for each case. Usually these are given as 1 or 0, but they might be category names like "yes" and "no." Often we just know the proportions of successes,  $\hat{p}_1$  and  $\hat{p}_2$ , and the counts,  $n_1$  and  $n_2$ . Computer packages don't usually deal with summary data like these easily. Calculators typically do a better job.

## EXERCISES

1. **Online social networking.** The Parents & Teens 2006 Survey of 935 12- to 17-year-olds found that, among teens aged 15–17, girls were significantly more likely to have used social networking sites and online profiles. 70% of the girls surveyed had used an online social network, compared to 54% of the boys. What does it mean to say that the difference in proportions is "significant"?
2. **Science news.** In 2007 a Pew survey asked 1447 Internet users about their sources of news and information about science. Among those who had broadband access at home, 34% said they would turn to the Internet for most of their science news. The report on this survey claims that this is not significantly different from the percentage (33%) who said they ordinarily get their science news from television. What does it mean to say that the difference is not significant?
3. **Name recognition.** A political candidate runs a week-long series of TV ads designed to attract public attention to his campaign. Polls taken before and after the ad campaign show some increase in the proportion of voters who now recognize this candidate's name, with a P-value of 0.033. Is it reasonable to believe the ads may be effective?
4. **Origins.** In a 1993 Gallup poll, 47% of the respondents agreed with the statement "*God created human beings pretty much in their present form at one time within the last 10,000 years or so.*" When Gallup asked the same question in 2001, only 45% of those respondents agreed. Is it reasonable to conclude that there was a change in public opinion given that the P-value is 0.37? Explain.
5. **Revealing information.** 886 randomly sampled teens were asked which of several personal items of information they thought it okay to share with someone they had just met. 44% said it was okay to share their e-mail addresses, but only 29% said they would give out their cell phone numbers. A researcher claims that a two-proportion z-test could tell whether there was a real difference among all teens. Explain why that test would not be appropriate for these data.
6. **Regulating access.** When a random sample of 935 parents were asked about rules in their homes, 77% said they had rules about the kinds of TV shows their children could watch. Among the 790 of those parents whose teenage children had Internet access, 85% had rules about the kinds of Internet sites their teens could visit. That looks like a difference, but can we tell? Explain why a two-sample z-test would not be appropriate here.
7. **Gender gap.** A presidential candidate fears he has a problem with women voters. His campaign staff plans to run a poll to assess the situation. They'll randomly sample 300 men and 300 women, asking if they have a favorable impression of the candidate. Obviously, the staff can't know this, but suppose the candidate has a positive image with 59% of males but with only 53% of females.
  - a) What sampling design is his staff planning to use?
  - b) What difference would you expect the poll to show?
  - c) Of course, sampling error means the poll won't reflect the difference perfectly. What's the standard deviation for the difference in the proportions?
  - d) Sketch a sampling model for the size difference in proportions of men and women with favorable impressions of this candidate that might appear in a poll like this.
  - e) Could the campaign be misled by the poll, concluding that there really is no gender gap? Explain.
8. **Buy it again?** A consumer magazine plans to poll car owners to see if they are happy enough with their vehicles that they would purchase the same model again. They'll randomly select 450 owners of American-made cars and 450 owners of Japanese models. Obviously, the actual opinions of the entire population couldn't be

- known, but suppose 76% of owners of American cars and 78% of owners of Japanese cars would purchase another.
- What sampling design is the magazine planning to use?
  - What difference would you expect their poll to show?
  - Of course, sampling error means the poll won't reflect the difference perfectly. What's the standard deviation for the difference in the proportions?
  - Sketch a sampling model for the difference in proportions that might appear in a poll like this.
  - Could the magazine be misled by the poll, concluding that owners of American cars are much happier with their vehicles than owners of Japanese cars? Explain.
9. **Arthritis.** The Centers for Disease Control and Prevention reported a survey of randomly selected Americans age 65 and older, which found that 411 of 1012 men and 535 of 1062 women suffered from some form of arthritis.
- Are the assumptions and conditions necessary for inference satisfied? Explain.
  - Create a 95% confidence interval for the difference in the proportions of senior men and women who have this disease.
  - Interpret your interval in this context.
  - Does this confidence interval suggest that arthritis is more likely to afflict women than men? Explain.
10. **Graduation.** In October 2000 the U.S. Department of Commerce reported the results of a large-scale survey on high school graduation. Researchers contacted more than 25,000 Americans aged 24 years to see if they had finished high school; 84.9% of the 12,460 males and 88.1% of the 12,678 females indicated that they had high school diplomas.
- Are the assumptions and conditions necessary for inference satisfied? Explain.
  - Create a 95% confidence interval for the difference in graduation rates between males and females.
  - Interpret your confidence interval.
  - Does this provide strong evidence that girls are more likely than boys to complete high school? Explain.
11. **Pets.** Researchers at the National Cancer Institute released the results of a study that investigated the effect of weed-killing herbicides on house pets. They examined 827 dogs from homes where an herbicide was used on a regular basis, diagnosing malignant lymphoma in 473 of them. Of the 130 dogs from homes where no herbicides were used, only 19 were found to have lymphoma.
- What's the standard error of the difference in the two proportions?
  - Construct a 95% confidence interval for this difference.
  - State an appropriate conclusion.
12. **Carpal tunnel.** The painful wrist condition called carpal tunnel syndrome can be treated with surgery or less invasive wrist splints. In September 2002, *Time* magazine reported on a study of 176 patients. Among the half that had surgery, 80% showed improvement after three months, but only 54% of those who used the wrist splints improved.
- What's the standard error of the difference in the two proportions?
  - Construct a 95% confidence interval for this difference.
  - State an appropriate conclusion.
13. **Ear infections.** A new vaccine was recently tested to see if it could prevent the painful and recurrent ear infections that many infants suffer from. *The Lancet*, a medical journal, reported a study in which babies about a year old were randomly divided into two groups. One group received vaccinations; the other did not. During the following year, only 333 of 2455 vaccinated children had ear infections, compared to 499 of 2452 unvaccinated children in the control group.
- Are the conditions for inference satisfied?
  - Find a 95% confidence interval for the difference in rates of ear infection.
  - Use your confidence interval to explain whether you think the vaccine is effective.
14. **Anorexia.** The *Journal of the American Medical Association* reported on an experiment intended to see if the drug Prozac<sup>®</sup> could be used as a treatment for the eating disorder anorexia nervosa. The subjects, women being treated for anorexia, were randomly divided into two groups. Of the 49 who received Prozac, 35 were deemed healthy a year later, compared to 32 of the 44 who got the placebo.
- Are the conditions for inference satisfied?
  - Find a 95% confidence interval for the difference in outcomes.
  - Use your confidence interval to explain whether you think Prozac is effective.
15. **Another ear infection.** In Exercise 13 you used a confidence interval to examine the effectiveness of a vaccine against ear infections in babies. Suppose that instead you had conducted a hypothesis test. (Answer these questions *without* actually doing the test.)
- What hypotheses would you test?
  - State a conclusion based on your confidence interval.
  - What alpha level did your test use?
  - If that conclusion is wrong, which type of error did you make?
  - What would be the consequences of such an error?
16. **Anorexia again.** In Exercise 14 you used a confidence interval to examine the effectiveness of Prozac in treating anorexia nervosa. Suppose that instead you had conducted a hypothesis test. (Answer these questions *without* actually doing the test.)
- What hypotheses would you test?
  - State a conclusion based on your confidence interval.
  - What alpha level did your test use?
  - If that conclusion is wrong, which type of error did you make?
  - What would be the consequences of such an error?
17. **Teen smoking, part I.** A Vermont study published in December 2001 by the American Academy of Pediatrics examined parental influence on teenagers' decisions to smoke. A group of students who had never smoked were questioned about their parents' attitudes toward smoking. These students were questioned again two years later to see if they had started smoking. The researchers found that, among the 284 students who indicated that their parents disapproved of kids smoking, 54 had become established smokers. Among the 41 students who initially said their parents were lenient about smoking, 11 became



smokers. Do these data provide strong evidence that parental attitude influences teenagers' decisions about smoking?

- a) What kind of design did the researchers use?
  - b) Write appropriate hypotheses.
  - c) Are the assumptions and conditions necessary for inference satisfied?
  - d) Test the hypothesis and state your conclusion.
  - e) Explain in this context what your P-value means.
  - f) If that conclusion is actually wrong, which type of error did you commit?
18. **Depression.** A study published in the *Archives of General Psychiatry* in March 2001 examined the impact of depression on a patient's ability to survive cardiac disease. Researchers identified 450 people with cardiac disease, evaluated them for depression, and followed the group for 4 years. Of the 361 patients with no depression, 67 died. Of the 89 patients with minor or major depression, 26 died. Among people who suffer from cardiac disease, are depressed patients more likely to die than non-depressed ones?
- a) What kind of design was used to collect these data?
  - b) Write appropriate hypotheses.
  - c) Are the assumptions and conditions necessary for inference satisfied?
  - d) Test the hypothesis and state your conclusion.
  - e) Explain in this context what your P-value means.
  - f) If your conclusion is actually incorrect, which type of error did you commit?
19. **Teen smoking, part II.** Consider again the Vermont study discussed in Exercise 17.
- a) Create a 95% confidence interval for the difference in the proportion of children who may smoke and have approving parents and those who may smoke and have disapproving parents.
  - b) Interpret your interval in this context.
  - c) Carefully explain what "95% confidence" means.
20. **Depression revisited.** Consider again the study of the association between depression and cardiac disease survivability in Exercise 18.
- a) Create a 95% confidence interval for the difference in survival rates.
  - b) Interpret your interval in this context.
  - c) Carefully explain what "95% confidence" means.
21. **Pregnancy.** In 1998, a San Diego reproductive clinic reported 42 live births to 157 women under the age of 38, but only 7 live births for 89 clients aged 38 and older. Is this strong evidence of a difference in the effectiveness of the clinic's methods for older women?
- a) Was this an experiment? Explain.
  - b) Test an appropriate hypothesis and state your conclusion in context.
  - c) If you concluded there was a difference, estimate that difference with a confidence interval and interpret your interval in context.
22. **Birthweight.** In 2003 the *Journal of the American Medical Association* reported a study examining the possible impact of air pollution caused by the 9/11 attack on New York's World Trade Center on the weight of babies.
- Researchers found that 8% of 182 babies born to mothers who were exposed to heavy doses of soot and ash on September 11 were classified as having low birth weight. Only 4% of 2300 babies born in another New York City hospital whose mothers had not been near the site of the disaster were similarly classified. Does this indicate a possibility that air pollution might be linked to a significantly higher proportion of low-weight babies?
- a) Was this an experiment? Explain.
  - b) Test an appropriate hypothesis and state your conclusion in context.
  - c) If you concluded there is a difference, estimate that difference with a confidence interval and interpret that interval in context.
23. **Politics and sex.** One month before the election, a poll of 630 randomly selected voters showed 54% planning to vote for a certain candidate. A week later it became known that he had had an extramarital affair, and a new poll showed only 51% of 1010 voters supporting him. Do these results indicate a decrease in voter support for his candidacy?
- a) Test an appropriate hypothesis and state your conclusion.
  - b) If your conclusion turns out to be wrong, did you make a Type I or Type II error?
  - c) If you concluded there was a difference, estimate that difference with a confidence interval and interpret your interval in context.
24. **Shopping.** A survey of 430 randomly chosen adults found that 21% of the 222 men and 18% of the 208 women had purchased books online.
- a) Is there evidence that men are more likely than women to make online purchases of books? Test an appropriate hypothesis and state your conclusion in context.
  - b) If your conclusion in fact proves to be wrong, did you make a Type I or Type II error?
  - c) Estimate this difference with a confidence interval.
  - d) Interpret your interval in context.
25. **Twins.** In 2001, one county reported that, among 3132 white women who had babies, 94 were multiple births. There were also 20 multiple births to 606 black women. Does this indicate any racial difference in the likelihood of multiple births?
- a) Test an appropriate hypothesis and state your conclusion in context.
  - b) If your conclusion is incorrect, which type of error did you commit?
26. **Mammograms.** A 9-year study in Sweden compared 21,088 women who had mammograms with 21,195 who did not. Of the women who underwent screening, 63 died of breast cancer, compared to 66 deaths among the control group. (*The New York Times*, Dec 9, 2001)
- a) Do these results support the effectiveness of regular mammograms in preventing deaths from breast cancer?
  - b) If your conclusion is incorrect, what kind of error have you committed?
27. **Pain.** Researchers comparing the effectiveness of two pain medications randomly selected a group of patients

who had been complaining of a certain kind of joint pain. They randomly divided these people into two groups, then administered the pain killers. Of the 112 people in the group who received medication A, 84 said this pain reliever was effective. Of the 108 people in the other group, 66 reported that pain reliever B was effective.

- Write a 95% confidence interval for the percent of people who may get relief from this kind of joint pain by using medication A. Interpret your interval.
  - Write a 95% confidence interval for the percent of people who may get relief by using medication B. Interpret your interval.
  - Do the intervals for A and B overlap? What do you think this means about the comparative effectiveness of these medications?
  - Find a 95% confidence interval for the difference in the proportions of people who may find these medications effective. Interpret your interval.
  - Does this interval contain zero? What does that mean?
  - Why do the results in parts c and e seem contradictory? If we want to compare the effectiveness of these two pain relievers, which is the correct approach? Why?
28. **Gender gap.** Candidates for political office realize that different levels of support among men and women may be a crucial factor in determining the outcome of an election. One candidate finds that 52% of 473 men polled say they will vote for him, but only 45% of the 522 women in the poll express support.
- Write a 95% confidence interval for the percent of male voters who may vote for this candidate. Interpret your interval.
  - Write and interpret a 95% confidence interval for the percent of female voters who may vote for him.
  - Do the intervals for males and females overlap? What do you think this means about the gender gap?
  - Find a 95% confidence interval for the difference in the proportions of males and females who will vote for this candidate. Interpret your interval.
  - Does this interval contain zero? What does that mean?
  - Why do the results in parts c and e seem contradictory? If we want to see if there is a gender gap among voters with respect to this candidate, which is the correct approach? Why?
29. **Sensitive men.** In August 2004, *Time* magazine, reporting on a survey of men's attitudes, noted that "Young men are more comfortable than older men talking about their problems." The survey reported that 80 of 129 surveyed 18- to 24-year-old men and 98 of 184 25- to 34-year-old men said they were comfortable. What do you think? Is *Time's* interpretation justified by these numbers?

30. **Retention rates.** In 2004 the testing company ACT, Inc., reported on the percentage of first-year students at 4-year colleges who return for a second year. Their sample of 1139 students in private colleges showed a 74.9% retention rate, while the rate was 71.9% for the sample of 505 students at public colleges. Does this provide evidence that there's a difference in retention rates of first-year students at public and private colleges?
31. **Online activity checks.** Are more parents checking up on their teen's online activities? A Pew survey in 2004 found that 33% of 868 randomly sampled teens said that their parents checked to see what Web sites they visited. In 2006 the same question posed to 811 teens found 41% reporting such checks. Do these results provide evidence that more parents are checking?
32. **Computer gaming.** Who plays online or electronic games? A survey in 2006 found that 69% of 223 boys aged 12–14 said they "played computer or console games like Xbox or PlayStation . . . or games online." Of 248 boys aged 15–17, only 62% played these games. Is this evidence of a real age-based difference?



### JUST CHECKING Answers

- We're 90% confident that if members are contacted by e-mail, the donation rate will be between 4 and 18 percentage points higher than if they received regular mail.
- Since a difference of 0 is not in the confidence interval, we'd reject the null hypothesis. There is evidence that more members will donate if contacted by e-mail.
- The proportion from the sample in 1995 has variability, too. If we do a one-proportion z-test, we won't take that variability into account and our P-value will be incorrect.
- The difference in the proportions between the two years has more variability than either individual proportion. The variance of the difference is the sum of the two variances.

## REVIEW OF PART V

## From the Data at Hand to the World at Large

## Quick Review

What do samples really tell us about the populations from which they are drawn? Are the results of an experiment meaningful, or are they just sampling error? Statistical inference based on our understanding of sampling models can help answer these questions. Here's a brief summary of the key concepts and skills:

- ▶ Sampling models describe the variability of sample statistics using a remarkable result called the Central Limit Theorem.
  - When the number of trials is sufficiently large, proportions found in different samples vary according to an approximately Normal model.
  - When samples are sufficiently large, the means of different samples vary, with an approximately Normal model.
  - The variability of sample statistics decreases as sample size increases.
  - Statistical inference procedures are based on the Central Limit Theorem.
  - No inference procedure is valid unless the underlying assumptions are true. Always check the conditions before proceeding.
- ▶ A confidence interval uses a sample statistic (such as a proportion) to estimate a range of plausible values for the parameter of a population model.
  - All confidence intervals involve an estimate of the parameter, a margin of error, and a level of confidence.
  - For confidence intervals based on a given sample, the greater the margin of error, the higher the confidence.
  - At a given level of confidence, the larger the sample, the smaller the margin of error.
- ▶ A hypothesis test proposes a model for the population, then examines the observed statistics to see if that model is plausible.
  - A null hypothesis suggests a parameter value for the population model. Usually, we assume there is nothing interesting, unusual, or different about the sample results.
  - The alternative hypothesis states what we will believe if the sample results turn out to be inconsistent with our null model.
  - We compare the difference between the statistic and the hypothesized value with the standard deviation of the statistic. It's the sampling distribution of this ratio that gives us a P-value.
  - The P-value of the test is the conditional probability that the null model could produce results at least as extreme as those observed in the sample or the experiment just as a result of sampling error.
  - A low P-value indicates evidence against the null model. If it is sufficiently low, we reject the null model.
  - A high P-value indicates that the sample results are not inconsistent with the null model, so we cannot reject it. However, this does not prove the null model is true.
  - Sometimes we will mistakenly reject the null hypothesis even though it's actually true—that's called a Type I error. If we fail to reject a false null hypothesis, we commit a Type II error.
  - The power of a test measures its ability to detect a false null hypothesis.
  - You can lower the risk of a Type I error by requiring a higher standard of proof (lower P-value) before rejecting the null hypothesis. But this will raise the risk of a Type II error and decrease the power of the test.
  - The only way to increase the power of a test while decreasing the chance of committing either error is to design a study based on a larger sample.

And now for some opportunities to review these concepts and skills . . .

## REVIEW EXERCISES

1. **Herbal cancer.** A report in the *New England Journal of Medicine* (June 6, 2000) notes growing evidence that the herb *Aristolochia fangchi* can cause urinary tract cancer in those who take it. Suppose you are asked to design an experiment to study this claim. Imagine that you have data on urinary tract cancers in subjects who have used this herb and similar subjects who have not used it and that you can measure incidences of cancer and precancerous lesions in these subjects. State the null and alternative hypotheses you would use in your study.
2. **Colorblind.** Medical literature says that about 8% of males are colorblind. A university's introductory psychology course is taught in a large lecture hall. Among the students, there are 325 males. Each semester when the

professor discusses visual perception, he shows the class a test for colorblindness. The percentage of males who are colorblind varies from semester to semester.

- a) Is the sampling distribution model for the sample proportion likely to be Normal? Explain.
  - b) What are the mean and standard deviation of this sampling distribution model?
  - c) Sketch the sampling model, using the 68–95–99.7 Rule.
  - d) Write a few sentences explaining what the model says about this professor's class.
3. **Birth days.** During a 2-month period in 2002, 72 babies were born at the Tompkins Community Hospital in upstate New York. The table shows how many babies were born on each day of the week.
- | Day    | Births |
|--------|--------|
| Mon.   | 7      |
| Tues.  | 17     |
| Wed.   | 8      |
| Thurs. | 12     |
| Fri.   | 9      |
| Sat.   | 10     |
| Sun.   | 9      |
- a) If births are uniformly distributed across all days of the week, how many would you expect on each day?
  - b) Only 7 births occurred on a Monday. Does this indicate that women might be less likely to give birth on a Monday? Explain.
  - c) Are the 17 births on Tuesdays unusually high? Explain.
  - d) Can you think of any reasons why births may not occur completely at random?
4. **Polling 2004.** In the 2004 U.S. presidential election, the official results showed that George W. Bush received 50.7% of the vote and John Kerry received 48.3%. Ralph Nader, running as a third-party candidate, picked up only 0.4%. After the election, there was much discussion about exit polls, which had initially indicated a different result. Suppose you had taken a random sample of 1000 voters in an exit poll and asked them for whom they had voted.
- a) Would you always get 507 votes for Bush and 483 for Kerry?
  - b) In 95% of such polls, your sample proportion of voters for Bush should be between what two values?
  - c) In 95% of such polls, your sample proportion of voters for Nader should be between what two numbers?
  - d) Would you expect the sample proportion of Nader votes to vary more, less, or about the same as the sample proportion of Bush votes? Why?
5. **Leaky gas tanks.** Nationwide, it is estimated that 40% of service stations have gas tanks that leak to some extent. A new program in California is designed to lessen the prevalence of these leaks. We want to assess the effectiveness of the program by seeing if the percentage of service stations whose tanks leak has decreased. To do this, we randomly sample 27 service stations in California and determine whether there is any evidence of leakage. In our sample, only 7 of the stations exhibit any leakage. Is there evidence that the new program is effective?
- a) What are the null and alternative hypotheses?
  - b) Check the assumptions necessary for inference.
  - c) Test the null hypothesis.
  - d) What do you conclude (in plain English)?
  - e) If the program actually works, have you made an error? What kind?

- f) What two things could you do to decrease the probability of making this kind of error?
- g) What are the advantages and disadvantages of taking those two courses of action?

6. **Surgery and germs.** Joseph Lister (for whom Listerine is named!) was a British physician who was interested in the role of bacteria in human infections. He suspected that germs were involved in transmitting infection, so he tried using carbolic acid as an operating room disinfectant. In 75 amputations, he used carbolic acid 40 times. Of the 40 amputations using carbolic acid, 34 of the patients lived. Of the 35 amputations without carbolic acid, 19 patients lived. The question of interest is whether carbolic acid is effective in increasing the chances of surviving an amputation.
- a) What kind of a study is this?
  - b) What do you conclude? Support your conclusion by testing an appropriate hypothesis.
  - c) What reservations do you have about the design of the study?
7. **Scrabble.** Using a computer to play many simulated games of Scrabble, researcher Charles Robinove found that the letter "A" occurred in 54% of the hands. This study had a margin of error of  $\pm 10\%$ . (*Chance*, 15, no. 1 [2002])
- a) Explain what the margin of error means in this context.
  - b) Why might the margin of error be so large?
  - c) Probability theory predicts that the letter "A" should appear in 63% of the hands. Does this make you concerned that the simulation might be faulty? Explain.
8. **Dice.** When one die is rolled, the number of spots showing has a mean of 3.5 and a standard deviation of 1.7. Suppose you roll 10 dice. What's the approximate probability that your total is between 30 and 40 (that is, the average for the 10 dice is between 3 and 4)? Specify the model you use and the assumptions and conditions that justify your approach.
9. **News sources.** In May of 2000, the Pew Research Foundation sampled 1593 respondents and asked how they obtain news. In Pew's report, 33% of respondents say that they now obtain news from the Internet at least once a week.
- a) Pew reports a margin of error of  $\pm 3\%$  for this result. Explain what the margin of error means.
  - b) Pew also asked about investment information, and 21% of respondents reported that the Internet is their main source of this information. When limited to the 780 respondents who identified themselves as investors, the percent who rely on the Internet rose to 28%. How would you expect the margin of error for this statistic to change in comparison with the margin of error for the percentage of all respondents?
  - c) When restricted to the 239 active traders in the sample, Pew reports that 45% rely on the Internet for investment information. Find a confidence interval for this statistic.
  - d) How does the margin of error for your confidence interval compare with the values in parts a and b? Explain why.

10. **Death penalty 2006.** In May of 2006, the Gallup Organization asked a random sample of 537 American adults this question:

*If you could choose between the following two approaches, which do you think is the better penalty for murder, the death penalty or life imprisonment, with absolutely no possibility of parole?*

Of those polled, 47% chose the death penalty, the lowest percentage in the 21 years that Gallup has asked this question.

- Create a 95% confidence interval for the percentage of all American adults who favor the death penalty.
  - Based on your confidence interval, is it clear that the death penalty no longer has majority support? Explain.
  - If pollsters wanted to follow up on this poll with another survey that could determine the level of support for the death penalty to within 2% with 98% confidence, how many people should they poll?
11. **Bimodal.** We are sampling randomly from a distribution known to be bimodal.
- As our sample size increases, what's the expected shape of the sample's distribution?
  - What's the expected value of our sample's mean? Does the size of the sample matter?
  - How is the variability of sample means related to the standard deviation of the population? Does the size of the sample matter?
  - How is the shape of the sampling distribution model affected by the sample size?
12. **Vitamin D.** In July 2002 the *American Journal of Clinical Nutrition* reported that 42% of 1546 African-American women studied had vitamin D deficiency. The data came from a national nutrition study conducted by the Centers for Disease Control and Prevention in Atlanta.
- Do these data meet the assumptions necessary for inference? What would you like to know that you don't?
  - Create a 95% confidence interval.
  - Interpret the interval in this context.
  - Explain in this context what "95% confidence" means.
13. **Archery.** A champion archer can generally hit the bull's-eye 80% of the time. Suppose she shoots 200 arrows during competition. Let  $\hat{p}$  represent the percentage of bull's-eyes she gets (the sample proportion).
- What are the mean and standard deviation of the sampling distribution model for  $\hat{p}$ ?
  - Is a Normal model appropriate here? Explain.
  - Sketch the sampling model, using the 68–95–99.7 Rule.
  - What's the probability that she gets at least 85% bull's-eyes?
14. **Free throws 2007.** During the 2006–2007 NBA season, Kyle Korver led the league by making 191 of 209 free throws, for a success rate of 91.39%. But Matt Carroll was close behind, with 188 of 208 (90.39%).
- Find a 95% confidence interval for the difference in their free throw percentages.
  - Based on your confidence interval, is it certain that Korver is better than Carroll at making free throws?
15. **Twins.** There is some indication in medical literature that doctors may have become more aggressive in inducing labor or doing preterm cesarean sections when a woman is carrying twins. Records at a large hospital show that, of the 43 sets of twins born in 1990, 20 were delivered before the 37th week of pregnancy. In 2000, 26 of 48 sets of twins were born preterm. Does this indicate an increase in the incidence of early births of twins? Test an appropriate hypothesis and state your conclusion.
16. **Eclampsia.** It's estimated that 50,000 pregnant women worldwide die each year of eclampsia, a condition involving elevated blood pressure and seizures. A research team from 175 hospitals in 33 countries investigated the effectiveness of magnesium sulfate in preventing the occurrence of eclampsia in at-risk patients. Results are summarized below. (*Lancet*, June 1, 2002)

	Total Subjects	Reported side effects	Developed eclampsia	Deaths
Treatment				
Magnesium sulfate	4999	1201	40	11
Placebo	4993	228	96	20

- Write a 95% confidence interval for the increase in the proportion of women who may develop side effects from this treatment. Interpret your interval.
- Is there evidence that the treatment may be effective in preventing the development of eclampsia? Test an appropriate hypothesis and state your conclusion.

17. **Eclampsia.** Refer again to the research summarized in Exercise 16. Is there any evidence that when eclampsia does occur, the magnesium sulfide treatment may help prevent the woman's death?
- Write an appropriate hypothesis.
  - Check the assumptions and conditions.
  - Find the P-value of the test.
  - What do you conclude about the magnesium sulfide treatment?
  - If your conclusion is wrong, which type of error have you made?
  - Name two things you could do to increase the power of this test.
  - What are the advantages and disadvantages of those two options?
18. **Eggs.** The ISA Babcock Company supplies poultry farmers with hens, advertising that a mature B300 Layer produces eggs with a mean weight of 60.7 grams. Suppose that egg weights follow a Normal model with standard deviation 3.1 grams.
- What fraction of the eggs produced by these hens weigh more than 62 grams?
  - What's the probability that a dozen randomly selected eggs average more than 62 grams?
  - Using the 68–95–99.7 Rule, sketch a model of the total weights of a dozen eggs.

19. **Polling disclaimer.** A newspaper article that reported the results of an election poll included the following explanation:

*The Associated Press poll on the 2000 presidential campaign is based on telephone interviews with 798 randomly selected registered voters from all states except Alaska and Hawaii. The interviews were conducted June 21–25 by ICR of Media, Pa.*

*The results were weighted to represent the population by demographic factors such as age, sex, region, and education.*

*No more than 1 time in 20 should chance variations in the sample cause the results to vary by more than 4 percentage points from the answers that would be obtained if all Americans were polled.*

*The margin of sampling error is larger for responses of subgroups, such as income categories or those in political parties. There are other sources of potential error in polls, including the wording and order of questions.*

- Did they describe the 5 W's well?
  - What kind of sampling design could take into account the several demographic factors listed?
  - What was the margin of error of this poll?
  - What was the confidence level?
  - Why is the margin of error larger for subgroups?
  - Which kinds of potential bias did they caution readers about?
20. **Enough eggs?** One of the important issues for poultry farmers is the production rate—the percentage of days on which a given hen actually lays an egg. Ideally, that would be 100% (an egg every day), but realistically, hens tend to lay eggs on about 3 of every 4 days. ISA Babcock wants to advertise the production rate for the B300 Layer (see Exercise 18) as a 95% confidence interval with a margin of error of  $\pm 2\%$ . How many hens must they collect data on?
21. **Teen deaths.** Traffic accidents are the leading cause of death among people aged 15 to 20. In May 2002, the National Highway Traffic Safety Administration reported that even though only 6.8% of licensed drivers are between 15 and 20 years old, they were involved in 14.3% of all fatal crashes. Insurance companies have long known that teenage boys were high risks, but what about teenage girls? One insurance company found that the driver was a teenage girl in 44 of the 388 fatal accidents they investigated. Is this strong evidence that the accident rate is lower for girls than for teens in general?
- Test an appropriate hypothesis and state your conclusion.
  - Explain what your P-value means in this context.
22. **Perfect pitch.** A recent study of perfect pitch tested students in American music conservatories. It found that 7% of 1700 non-Asian and 32% of 1000 Asian students have perfect pitch. A test of the difference in proportions resulted in a P-value of  $< 0.0001$ .
- What are the researchers' null and alternative hypotheses?
  - State your conclusion.
  - Explain in this context what the P-value means.
  - The researchers claimed that the data prove that genetic differences between the two populations cause a difference in the frequency of occurrence of perfect pitch. Do you agree? Why or why not?
23. **Largemouth bass.** Organizers of a fishing tournament believe that the lake holds a sizable population of largemouth bass. They assume that the weights of these fish have a model that is skewed to the right with a mean of 3.5 pounds and a standard deviation of 2.2 pounds.
- Explain why a skewed model makes sense here.
  - Explain why you cannot determine the probability that a largemouth bass randomly selected ("caught") from the lake weighs over 3 pounds.
  - Each fisherman in the contest catches 5 fish each day. Can you determine the probability that someone's catch averages over 3 pounds? Explain.
  - The 12 fishermen competing each caught the limit of 5 fish. What's the probability that the total catch of 60 fish averaged more than 3 pounds?
24. **Cheating.** A Rutgers University study released in 2002 found that many high school students cheat on tests. The researchers surveyed a random sample of 4500 high school students nationwide; 74% of them said they had cheated at least once.
- Create a 90% confidence interval for the level of cheating among high school students. Don't forget to check the appropriate conditions.
  - Interpret your interval.
  - Explain what "90% confidence" means.
  - Would a 95% confidence interval be wider or narrower? Explain without actually calculating the interval.
25. **Language.** Neurological research has shown that in about 80% of people language abilities reside in the brain's left side. Another 10% display right-brain language centers, and the remaining 10% have two-sided language control. (The latter two groups are mainly left-handers.) (Science News, 161, no. 24 [2002])
- We select 60 people at random. Is it reasonable to use a Normal model to describe the possible distribution of the proportion of the group that has left-brain language control? Explain.
  - What's the probability that our group has at least 75% left-brainers?
  - If the group had consisted of 100 people, would that probability be higher, lower, or about the same? Explain why, without actually calculating the probability.
  - How large a group would almost certainly guarantee at least 75% left-brainers? Explain.
26. **Cigarettes 2006.** In 1999 the Centers for Disease Control and Prevention estimated that about 34.8% of high school students smoked cigarettes. They established a national health goal of reducing that figure to 16% by the year 2010. To that end, they hoped to achieve a reduction to 20% by 2006. In 2006 they released a research study in which 23% of a random sample of 1815 high school students said they were current smokers. Is this evidence that progress toward the goal is off track?
- Write appropriate hypotheses.
  - Verify that the appropriate assumptions are satisfied.

- c) Find the P-value of this test.  
 d) Explain what the P-value means in this context.  
 e) State an appropriate conclusion.  
 f) Of course, your conclusion may be incorrect. If so, which kind of error did you commit?
27. **Crohn's disease.** In 2002 the medical journal *The Lancet* reported that 335 of 573 patients suffering from Crohn's disease responded positively to injections of the arthritis-fighting drug infliximab.  
 a) Create a 95% confidence interval for the effectiveness of this drug.  
 b) Interpret your interval in context.  
 c) Explain carefully what "95% confidence" means in this context.
28. **Teen smoking 2006.** The Centers for Disease Control and Prevention say that about 23% of teenagers smoke tobacco (down from a high of 38% in 1997). A college has 522 students in its freshman class. Is it likely that more than 30% of them are smokers? Explain.
29. **Alcohol abuse.** Growing concern about binge drinking among college students has prompted one large state university to conduct a survey to assess the size of the problem on its campus. The university plans to randomly select students and ask how many have been drunk during the past week. If the school hopes to estimate the true proportion among all its students with 90% confidence and a margin of error of  $\pm 4\%$ , how many students must be surveyed?
30. **Errors.** An auto parts company advertises that its special oil additive will make the engine "run smoother, cleaner, longer, with fewer repairs." An independent laboratory decides to test part of this claim. It arranges to use a taxicab company's fleet of cars. The cars are randomly divided into two groups. The company's mechanics will use the additive in one group of cars but not in the other. At the end of a year the laboratory will compare the percentage of cars in each group that required engine repairs.  
 a) What kind of a study is this?  
 b) Will they do a one-tailed or a two-tailed test?  
 c) Explain in this context what a Type I error would be.  
 d) Explain in this context what a Type II error would be.  
 e) Which type of error would the additive manufacturer consider more serious?  
 f) If the cabs with the additive do indeed run significantly better, can the company conclude it is an effect of the additive? Can they generalize this result and recommend the additive for all cars? Explain.
31. **Preemies.** Among 242 Cleveland-area children born prematurely at low birth weights between 1977 and 1979, only 74% graduated from high school. Among a comparison group of 233 children of normal birth weight, 83% were high school graduates. ("Outcomes in Young Adulthood for Very-Low-Birth-Weight Infants," *New England Journal of Medicine*, 346, no. 3 [2002])  
 a) Create a 95% confidence interval for the difference in graduation rates between children of normal and children of very low birth weights. Be sure to check the appropriate assumptions and conditions.  
 b) Does this provide evidence that premature birth may be a risk factor for not finishing high school? Use your confidence interval to test an appropriate hypothesis.  
 c) Suppose your conclusion is incorrect. Which type of error did you make?
32. **Safety.** Observers in Texas watched children at play in eight communities. Of the 814 children seen biking, roller skating, or skateboarding, only 14% wore a helmet.  
 a) Create and interpret a 95% confidence interval.  
 b) What concerns do you have about this study that might make your confidence interval unreliable?  
 c) Suppose we want to do this study again, picking various communities and locations at random, and hope to end up with a 98% confidence interval having a margin of error of  $\pm 4\%$ . How many children must we observe?
33. **Fried PCs.** A computer company recently experienced a disastrous fire that ruined some of its inventory. Unfortunately, during the panic of the fire, some of the damaged computers were sent to another warehouse, where they were mixed with undamaged computers. The engineer responsible for quality control would like to check out each computer in order to decide whether it's undamaged or damaged. Each computer undergoes a series of 100 tests. The number of tests it fails will be used to make the decision. If it fails more than a certain number, it will be classified as damaged and then scrapped. From past history, the distribution of the number of tests failed is known for both undamaged and damaged computers. The probabilities associated with each outcome are listed in the table below:

Number of tests failed	0	1	2	3	4	5	>5
Undamaged (%)	80	13	2	4	1	0	0
Damaged (%)	0	10	70	5	4	1	10

The table indicates, for example, that 80% of the undamaged computers have no failures, while 70% of the damaged computers have 2 failures.

- a) To the engineers, this is a hypothesis-testing situation. State the null and alternative hypotheses.  
 b) Someone suggests classifying a computer as damaged if it fails any of the tests. Discuss the advantages and disadvantages of this test plan.  
 c) What number of tests would a computer have to fail in order to be classified as damaged if the engineers want to have the probability of a Type I error equal to 5%?  
 d) What's the power of the test plan in part c?  
 e) A colleague points out that by increasing  $\alpha$  just 2%, the power can be increased substantially. Explain.
34. **Power.** We are replicating an experiment. How will each of the following changes affect the power of our test? Indicate whether it will increase, decrease, or remain the same, assuming that all other aspects of the situation remain unchanged.  
 a) We increase the number of subjects from 40 to 100.  
 b) We require a higher standard of proof, changing from  $\alpha = 0.05$  to  $\alpha = 0.01$ .

35. **Approval 2007.** Of all the post–World War II presidents, Richard Nixon had the highest *disapproval* rating near the end of his presidency. His disapproval rating peaked at 66% in July 1974, just before he resigned. In May 2007, George W. Bush’s disapproval rating was 63%, according to a Gallup poll of 1000 voters. Pundits started discussing whether his rating was still discernibly better than Nixon’s. What do you think?
36. **Grade inflation.** In 1996, 20% of the students at a major university had an overall grade point average of 3.5 or higher (on a scale of 4.0). In 2000, a random sample of 1100 student records found that 25% had a GPA of 3.5 or higher. Is this evidence of grade inflation?
37. **Name recognition.** An advertising agency won’t sign an athlete to do product endorsements unless it is sure the person is known to more than 25% of its target audience. The agency always conducts a poll of 500 people to investigate the athlete’s name recognition before offering a contract. Then it tests  $H_0: p = 0.25$  against  $H_A: p > 0.25$  at a 5% level of significance.
- Why does the company use upper tail tests in this situation?
  - Explain what Type I and Type II errors would represent in this context, and describe the risk that each error poses to the company.
  - The company is thinking of changing its test to use a 10% level of significance. How would this change the company’s exposure to each type of risk?
38. **Name recognition, part II.** The advertising company described in Exercise 37 is thinking about signing a WNBA star to an endorsement deal. In its poll, 27% of the respondents could identify her.
- Fans who never took Statistics can’t understand why the company did not offer this WNBA player an endorsement contract even though the 27% recognition rate in the poll is above the 25% threshold. Explain it to them.
  - Suppose that further polling reveals that this WNBA star really is known to about 30% of the target audience. Did the company initially commit a Type I or Type II error in not signing her?
  - Would the power of the company’s test have been higher or lower if the player were more famous? Explain.
39. **NIMBY.** In March 2007, the Gallup Poll split a sample of 1003 randomly selected U.S. adults into two groups at random. Half ( $n = 502$ ) of the respondents were asked,
- “Overall, do you strongly favor, somewhat favor, somewhat oppose, or strongly oppose the use of nuclear energy as one of the ways to provide electricity for the U.S.?”*
- They found that 53% were either “somewhat” or “strongly” in favor. The other half ( $n = 501$ ) were asked,
- “Overall, would you strongly favor, somewhat favor, somewhat oppose, or strongly oppose the construction of a nuclear energy plant in your area as one of the ways to provide electricity for the U.S.?”*
- Only 40% were somewhat or strongly in favor. This difference is an example of the NIMBY (Not In My Back-Yard) phenomenon and is a serious concern to policy makers and planners. How large is the difference between the proportion of American adults who think nuclear energy is a good idea and the proportion who would be willing to have a nuclear plant in their area? Construct and interpret an appropriate confidence interval.
40. **Dropouts.** One study comparing various treatments for the eating disorder anorexia nervosa initially enlisted 198 subjects, but found overall that 105 failed to complete their assigned treatment programs. Construct and interpret an appropriate confidence interval. Discuss any reservations you have about this inference.





PART

VI

# Learning About the World

## Chapter 23

Inferences About Means

## Chapter 24

Comparing Means

## Chapter 25

Paired Samples and Blocks

# Inferences About Means



<b>WHO</b>	Vehicles on Triphammer Road
<b>WHAT</b>	Speed
<b>UNITS</b>	Miles per hour
<b>WHEN</b>	April 11, 2000, 1 p.m.
<b>WHERE</b>	A small town in the northeastern United States
<b>WHY</b>	Concern over impact on residential neighborhood

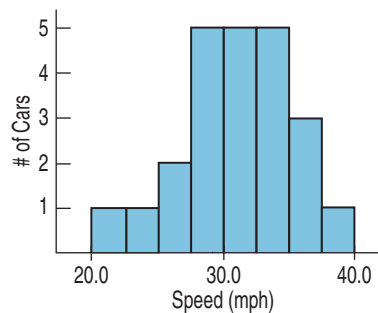
Motor vehicle crashes are the leading cause of death for people between 4 and 33 years old. In the year 2006, motor vehicle accidents claimed the lives of 43,300 people in the United States. This means that, on average, motor vehicle crashes resulted in 119 deaths each day, or 1 death every 12 minutes. Speeding is a contributing factor in 31% of all fatal accidents, according to the National Highway Traffic Safety Administration.

Triphammer Road is a busy street that passes through a residential neighborhood. Residents there are concerned that vehicles traveling on Triphammer often exceed the posted speed limit of 30 miles per hour. The local police sometimes place a radar speed detector by the side of the road; as a vehicle approaches, this detector displays the vehicle's speed to its driver.

The local residents are not convinced that such a passive method is helping the problem. They wish to persuade the village to add extra police patrols to encourage drivers to observe the speed limit. To help their case, a resident stood where he could see the detector and recorded the speed of vehicles passing it during a 15-minute period one day. When clusters of vehicles went by, he noted only the speed of the front vehicle. Here are his data and the histogram.

**FIGURE 23.1**

*The speeds of cars on Triphammer Road seem to be unimodal and symmetric, at least at this scale.*



Speed		
29	29	24
34	34	34
34	32	36
28	31	31
30	27	34
29	37	36
38	29	21
31	26	

We're interested both in estimating the true mean speed and in testing whether it exceeds the posted speed limit. Although the sample of vehicles is a convenience sample, not a truly random sample, there's no compelling reason to

believe that vehicles at one time of day are driving faster or slower than vehicles at another time of day,<sup>1</sup> so we can take the sample to be representative.

These data differ from data on proportions in one important way. Proportions are usually reported as summaries. After all, individual responses are just “success” and “failure” or “1” and “0.” Quantitative data, though, usually report a value for each individual. When you have a value for each individual, you should remember the three rules of data analysis and plot the data, as we have done here.

We have quantitative data, so we summarize with means and standard deviations. Because we want to make inferences, we’ll think about sampling distributions, too, and we already know most of the facts we need.

## Getting Started

You’ve learned how to create confidence intervals and test hypotheses about proportions. We always center confidence intervals at our best guess of the unknown parameter. Then we add and subtract a margin of error. For proportions, that means  $\hat{p} \pm ME$ .

We found the margin of error as the product of the standard error,  $SE(\hat{p})$ , and a critical value,  $z^*$ , from the Normal table. So we had  $\hat{p} \pm z^*SE(\hat{p})$ .

We knew we could use  $z$  because the Central Limit Theorem told us (back in Chapter 18) that the sampling distribution model for proportions is Normal.

Now we want to do exactly the same thing for means, and fortunately, the Central Limit Theorem (still in Chapter 18) told us that the same Normal model works as the sampling distribution for means.

### THE CENTRAL LIMIT THEOREM

When a random sample is drawn from any population with mean  $\mu$  and standard deviation  $\sigma$ , its sample mean,  $\bar{y}$ , has a sampling distribution with the same *mean*  $\mu$  but whose *standard deviation* is  $\frac{\sigma}{\sqrt{n}}$  (and we write  $\sigma(\bar{y}) = SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ ).

No matter what population the random sample comes from, the *shape* of the sampling distribution is approximately Normal as long as the sample size is large enough. The larger the sample used, the more closely the Normal approximates the sampling distribution for the mean.

### FOR EXAMPLE

#### Using the CLT (as if we knew $\sigma$ )

Based on weighing thousands of animals, the American Angus Association reports that mature Angus cows have a mean weight of 1309 pounds with a standard deviation of 157 pounds. This result was based on a very large sample of animals from many herds over a period of 15 years, so let’s assume that these summaries are the population parameters and that the distribution of the weights was unimodal and reasonably symmetric.

**Question:** What does the CLT predict about the mean weight seen in random samples of 100 mature Angus cows?

(continued)

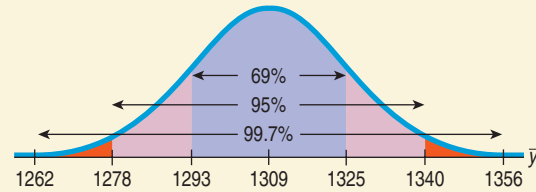
<sup>1</sup> Except, perhaps, at rush hour. But at that time, traffic is slowed. Our concern is with ordinary traffic during the day.

## For Example (continued)

It's given that weights of all mature Angus cows have  $\mu = 1309$  and  $\sigma = 157$  pounds. Because  $n = 100$  animals is a fairly large sample, I can apply the Central Limit Theorem. I expect the resulting sample means  $\bar{y}$  will average 1309 pounds and have a standard deviation of  $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{157}{\sqrt{100}} = 15.7$  pounds.

The CLT also says that the distribution of sample means follows a Normal model, so the 68–95–99.7 Rule applies. I'd expect that

- ▶ in 68% of random samples of 100 mature Angus cows, the mean weight will be between  $1309 - 15.7 = 1293.3$  and  $1309 + 15.7 = 1324.7$  pounds;
- ▶ in 95% of such samples,  $1277.6 \leq \bar{y} \leq 1340.4$  pounds;
- ▶ in 99.7% of such samples,  $1261.9 \leq \bar{y} \leq 1356.1$  pounds.



The CLT says that all we need to model the sampling distribution of  $\bar{y}$  is a random sample of quantitative data.

And the true population standard deviation,  $\sigma$ .

Uh oh. That could be a problem. How are we supposed to know  $\sigma$ ? With proportions, we had a link between the proportion value and the standard deviation of the sample proportion:  $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$ . And there was an obvious way to estimate

the standard deviation from the data:  $SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ . But for means,  $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ , so knowing  $\bar{y}$  doesn't tell us anything about  $SD(\bar{y})$ . We know  $n$ , the sample size, but the population standard deviation,  $\sigma$ , could be *anything*. So what should we do? We do what any sensible person would do: We estimate the population parameter  $\sigma$  with  $s$ , the sample standard deviation based on the data. The resulting standard error is  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ .

A century ago, people used this standard error with the Normal model, assuming it would work. And for large sample sizes it *did* work pretty well. But they began to notice problems with smaller samples. The sample standard deviation,  $s$ , like any other statistic, varies from sample to sample. And this extra variation in the standard error was messing up the P-values and margins of error.

William S. Gosset is the man who first investigated this fact. He realized that not only do we need to allow for the extra variation with larger margins of error and P-values, but we even need a new sampling distribution model. In fact, we need a whole *family* of models, depending on the sample size,  $n$ . These models are unimodal, symmetric, bell-shaped models, but the smaller our sample, the more we must stretch out the tails. Gosset's work transformed Statistics, but most people who use his work don't even know his name.

Because we estimate the standard deviation of the sampling distribution model from the data, it's a *standard error*. So we use the  $SE(\bar{y})$  notation. Remember, though, that it's just the estimated standard deviation of the sampling distribution model for means.

**AS** **Activity: Estimating the Standard Error.** What's the average age at which people have heart attacks? A confidence interval gives a good answer, but we must estimate the standard deviation from the data to construct the interval.

## Gosset's *t*

Gosset had a job that made him the envy of many. He was the quality control engineer for the Guinness Brewery in Dublin, Ireland. His job was to make sure that the stout (a thick, dark beer) leaving the brewery was of high enough quality to meet the demands of the brewery's many discerning customers. It's easy to imagine why a large sample with many observations might be undesirable when testing stout, not to mention dangerous to one's health. So Gosset often used small



To find the sampling distribution of  $\frac{\bar{y}}{s/\sqrt{n}}$ , Gosset simulated it by hand. He drew paper slips of small samples from a hat hundreds of times and computed the means and standard deviations with a mechanically cranked calculator. Today you could repeat in seconds on a computer the experiment that took him over a year. Gosset's work was so meticulous that not only did he get the shape of the new histogram approximately right, but he even figured out the exact formula for it from his sample. The formula was not confirmed mathematically until years later by Sir R. A. Fisher.

samples of 3 or 4. But he noticed that with samples of this size, his tests for quality weren't quite right. He knew this because when the batches that he rejected were sent back to the laboratory for more extensive testing, too often they turned out to be OK.

Gosset checked the stout's quality by performing hypothesis tests. He knew that the test would make some Type I errors and reject about 5% of the good batches of stout. However, the lab told him that he was in fact rejecting about 15% of the good batches. Gosset knew something was wrong, and it bugged him.

Gosset took time off to study the problem (and earn a graduate degree in the emerging field of Statistics). He figured out that when he used the standard error,  $\frac{s}{\sqrt{n}}$ , as an estimate of the standard deviation, the shape of the sampling model changed. He even figured out what the new model should be and called it a *t*-distribution.

The Guinness Company didn't give Gosset a lot of support for his work. In fact, it had a policy against publishing results. Gosset had to convince the company that he was not publishing an industrial secret, and (as part of getting permission to publish) he had to use a pseudonym. The pseudonym he chose was "Student," and ever since, the model he found has been known as **Student's *t***.

Gosset's model is always bell-shaped, but the details change with different sample sizes. So the Student's *t*-models form a whole family of related distributions that depend on a parameter known as **degrees of freedom**. We often denote degrees of freedom as *df* and the model as  $t_{df}$ , with the degrees of freedom as a subscript.

## A Confidence Interval for Means

To make confidence intervals or test hypotheses for means, we need to use Gosset's model. Which one? Well, for means, it turns out the right value for degrees of freedom is  $df = n - 1$ .

### NOTATION ALERT:

Ever since Gosset, *t* has been reserved in Statistics for his distribution.

### A PRACTICAL SAMPLING DISTRIBUTION MODEL FOR MEANS

When certain assumptions and conditions<sup>2</sup> are met, the standardized sample mean,

$$t = \frac{\bar{y} - \mu}{SE(\bar{y})},$$

follows a Student's *t*-model with  $n - 1$  degrees of freedom. We estimate the standard deviation with

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

When Gosset corrected the model for the extra uncertainty, the margin of error got bigger, as you might have guessed. When you use Gosset's model instead of the Normal model, your confidence intervals will be just a bit wider and your P-values just a bit larger. That's the correction you need. By using the *t*-model, you've compensated for the extra variability in precisely the right way.

<sup>2</sup> You can probably guess what they are. We'll see them in the next section.

**NOTATION ALERT:**

When we found critical values from a Normal model, we called them  $z^*$ . When we use a Student's  $t$ -model, we'll denote the critical values  $t^*$ .

**AS** **Activity: Student's  $t$  in Practice.** Use a statistics package to find a  $t$ -based confidence interval; that's how it's almost always done.

**ONE-SAMPLE  $t$ -INTERVAL FOR THE MEAN**

When the assumptions and conditions<sup>3</sup> are met, we are ready to find the confidence interval for the population mean,  $\mu$ . The confidence interval is

$$\bar{y} \pm t_{n-1}^* \times SE(\bar{y}),$$

where the standard error of the mean is  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ .

The critical value  $t_{n-1}^*$  depends on the particular confidence level,  $C$ , that you specify and on the number of degrees of freedom,  $n - 1$ , which we get from the sample size.

**FOR EXAMPLE****A one-sample  $t$ -interval for the mean**

In 2004, a team of researchers published a study of contaminants in farmed salmon.<sup>4</sup> Fish from many sources were analyzed for 14 organic contaminants. The study expressed concerns about the level of contaminants found. One of those was the insecticide mirex, which has been shown to be carcinogenic and is suspected to be toxic to the liver, kidneys, and endocrine system. One farm in particular produced salmon with very high levels of mirex. After those outliers are removed, summaries for the mirex concentrations (in parts per million) in the rest of the farmed salmon are:

$$n = 150 \quad \bar{y} = 0.0913 \text{ ppm} \quad s = 0.0495 \text{ ppm.}$$

**Question:** What does a 95% confidence interval say about mirex?

$$df = 150 - 1 = 149$$

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{0.0495}{\sqrt{150}} = 0.0040$$

$$t_{149}^* \approx 1.977 \text{ (from table } T, \text{ using } 140 \text{ } df)$$

$$\text{(actually, } t_{149}^* \approx 1.976 \text{ from technology)}$$

So the confidence interval for  $\mu$  is  $\bar{y} \pm t_{149}^* \times SE(\bar{y}) = 0.0913 \pm 1.977(0.0040)$

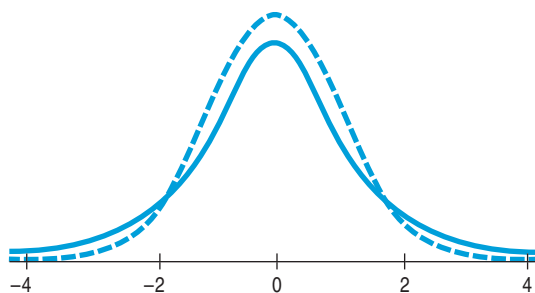
$$= 0.0913 \pm 0.0079$$

$$= (0.0834, 0.0992)$$

I'm 95% confident that the mean level of mirex concentration in farm-raised salmon is between 0.0834 and 0.0992 parts per million.

**FIGURE 23.2**

The  $t$ -model (solid curve) on 2 degrees of freedom has fatter tails than the Normal model (dashed curve). So the 68–95–99.7 Rule doesn't work for  $t$ -models with only a few degrees of freedom.



**AS** **Activity: Student's Distributions.** Interact with Gosset's family of  $t$ -models. Watch the shape of the model change as you slide the degrees of freedom up and down.

<sup>3</sup> Yes, the same ones, and they're still coming in the next section.

<sup>4</sup> Ronald A. Hites, Jeffery A. Foran, David O. Carpenter, M. Coreen Hamilton, Barbara A. Knuth, and Steven J. Schwager, "Global Assessment of Organic Contaminants in Farmed Salmon," *Science* 9 January 2004: Vol. 303., no. 5655, pp. 226–229.

TI-*n*spire

**The  $t$ -models.** See how  $t$ -models change as you change the degrees of freedom.

## z or t?

If you know  $\sigma$ , use  $z$ .  
(That's rare!)

Whenever you use  $\sigma$  to estimate  $\sigma$ , use  $t$ .

Student's  $t$ -models are unimodal, symmetric, and bell-shaped, just like the Normal. But  $t$ -models with only a few degrees of freedom have much fatter tails than the Normal. (That's what makes the margin of error bigger.) As the degrees of freedom increase, the  $t$ -models look more and more like the Normal. In fact, the  $t$ -model with infinite degrees of freedom is exactly Normal.<sup>5</sup> This is great news if you happen to have an infinite number of data values. Unfortunately, that's not practical. Fortunately, above a few hundred degrees of freedom it's very hard to tell the difference. Of course, in the rare situation that we *know*  $\sigma$ , it would be foolish not to use that information. And if we don't have to estimate  $\sigma$ , we can use the Normal model.

**When  $\sigma$  is known** Administrators of a hospital were concerned about the prenatal care given to mothers in their part of the city. To study this, they examined the gestation times of babies born there. They drew a sample of 25 babies born in their hospital in the previous 6 months. Human gestation times for healthy pregnancies are thought to be well-modeled by a Normal with a mean of 280 days and a standard deviation of 14 days. The hospital administrators wanted to test the mean gestation time of their sample of babies against the known standard. For this test, they should use the established value for the standard deviation, 14 days, rather than estimating the standard deviation from their sample. Because they use the model parameter value for  $\sigma$ , they should base their test on the Normal model rather than Student's  $t$ .

## TI Tips

Finding  $t$ -model probabilities and critical values

```
normalcdf(1.645,
99)
.0499848898
```

```
DISTR DRAW
1:normalPdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²Pdf(
```

```
.0499848898
tcdf(1.645,99,12
)
.0629457739
tcdf(1.645,99,25
)
.0562435022
```

## Finding Probabilities

You already know how to use your TI to find probabilities for Normal models using  $z$ -scores and `normalcdf`. What about  $t$ -models? Yes, the calculator can work with them, too.

You know from your experience with confidence intervals that  $z = 1.645$  cuts off the upper 5% in a Normal model. Use the TI to check that. From the **DISTR** menu, enter `normalcdf(1.645,99)`. Only 0.04998? Close enough for statisticians!

We might wonder about the probability of observing a  $t$ -value greater than 1.645, but we can't find that. There's only one Normal model, but there are many  $t$ -models, depending on the number of degrees of freedom. We need to be more specific.

Let's find the probability of observing a  $t$ -value greater than 1.645 when there are 12 degrees of freedom. That we can do. Look in the **DISTR** menu again. See it? Yes, `tcdf`. That function works essentially like `normalcdf`, but after you enter the left and right cutoffs you must also specify the number of degrees of freedom. Try `tcdf(1.645,99,12)`.

The upper tail probability for  $t_{12}$  is 0.063, higher than the Normal model's 0.05. That should make sense to you—remember,  $t$ -models are a bit fatter in the tails, so more of the distribution lies beyond the 1.645 cutoff. (That means we'll have to go a little wider to make a 90% confidence interval.)

<sup>5</sup> Formally, in the limit as  $n$  goes to infinity.

```

OSI: DRAW
1:normalpdf(
2:normalcdf(
3:invNorm(
4:invT(
5:tpdf(
6:tcdf(
7:χ²pdf(

```

```

invNorm(.99)
2.326347877
invT(.99,6)
3.142668396

```

Check out what happens when there are more degrees of freedom, say, 25. The command `tcdf(1.645,99,25)` yields a probability of 0.056. That's closer to 0.05, for a good reason:  $t$ -models look more and more like the Normal model as the number of degrees of freedom increases.

### Finding Critical Values

Your calculator can also determine the critical value of  $t$  that cuts off a specified percentage of the distribution, using `invT`. It works just like `invNorm`, but for  $t$  we also have to specify the number of degrees of freedom (of course).

Suppose we have 6 degrees of freedom and want to create a 98% confidence interval. A confidence level of 98% leaves 1% in each tail of our model, so we need to find the value of  $t$  corresponding to the 99th percentile. If a Normal model were appropriate, we'd use  $z = 2.33$ . (Try it: `invNorm(.99)`). Now think. How should the critical value for  $t$  compare?

If you thought, "It'll be larger, because  $t$ -models are more spread out," you're right. Check with your TI, remembering to specify our 6 degrees of freedom: `invT(.99,6)`. Were you surprised, though, that the critical value of  $t$  is so much larger?

So think once more. How would the critical value of  $t$  differ if there were 60 degrees of freedom instead of only 6? When you think you know, check it out on your TI.

### Understanding $t$

Use your calculator to play around with `tcdf` and `invT` a bit. Try to develop a clear understanding of how  $t$ -models compare to the more familiar Normal model. That will help you as you learn to use  $t$ -models to make inferences about means.

## Assumptions and Conditions

Gosset found the  $t$ -model by simulation. Years later, when Sir Ronald A. Fisher<sup>6</sup> showed mathematically that Gosset was right, he needed to make some assumptions to make it work. These are the assumptions we need to use the Student's  $t$ -models.

### INDEPENDENCE ASSUMPTION

**Independence Assumption:** The data values should be independent. There's really no way to check independence of the data by looking at the sample, but we should think about whether the assumption is reasonable.

**Randomization Condition:** The data arise from a random sample or suitably randomized experiment. Randomly sampled data—and especially data from a Simple Random Sample—are ideal.

When a sample is drawn without replacement, technically we ought to confirm that we haven't sampled a large fraction of the population, which would threaten the independence of our selections. We check the

**10% Condition:** The sample is no more than 10% of the population.

In practice, though, we often don't mention the 10% Condition for means. Why not? When we made inferences about proportions, this condition was crucial

<sup>6</sup> We met Fisher back in Chapter 21. You can see his picture on page 486.



### We Don't Want to Stop

We check conditions hoping that we can make a meaningful analysis of our data. The conditions serve as *disqualifiers*—we keep going unless there's a serious problem. If we find minor issues, we note them and express caution about our results. If the sample is not an SRS, but we believe it's representative of some populations, we limit our conclusions accordingly. If there are outliers, rather than stop, we perform the analysis both with and without them. If the sample looks bimodal, we try to analyze subgroups separately. Only when there's major trouble—like a strongly skewed small sample or an obviously nonrepresentative sample—are we unable to proceed at all.

because we usually had large samples. But for means our samples are generally smaller, so the independence problem arises only if we're sampling from a small population (and then there's a correction formula we could use—but let's not get into that here). And sometimes we're dealing with a randomized experiment; then there's no sampling at all.

## NORMAL POPULATION ASSUMPTION

Student's  $t$ -models won't work for data that are badly skewed. How skewed is too skewed? Well, formally, we assume that the data are from a population that follows a Normal model. Practically speaking, there's no way to be certain this is true.

And it's almost certainly *not* true. Models are idealized; real data are, well, real—*never* Normal. The good news, however, is that even for small samples, it's sufficient to check the . . .

**Nearly Normal Condition:** The data come from a distribution that is unimodal and symmetric.

Check this condition by making a histogram or Normal probability plot. The importance of Normality for Student's  $t$  depends on the sample size. Just our luck: It matters most when it's hardest to check.<sup>7</sup>

For very small samples ( $n < 15$  or so), the data should follow a Normal model pretty closely. Of course, with so little data, it's rather hard to tell. But if you do find outliers or strong skewness, don't use these methods.

For moderate sample sizes ( $n$  between 15 and 40 or so), the  $t$  methods will work well as long as the data are unimodal and reasonably symmetric. Make a histogram.

When the sample size is larger than 40 or 50, the  $t$  methods are safe to use unless the data are extremely skewed. Be sure to make a histogram. If you find outliers in the data, it's always a good idea to perform the analysis twice, once with and once without the outliers, even for large samples. They may well hold additional information about the data that deserves special attention. If you find multiple modes, you may well have different groups that should be analyzed and understood separately.

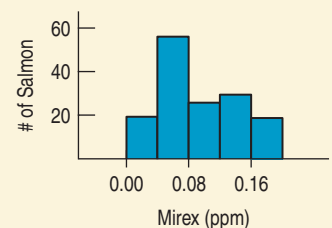
### FOR EXAMPLE

#### Checking assumptions and conditions for Student's $t$

**Recap:** Researchers purchased whole farmed salmon from 51 farms in eight regions in six countries. The histogram shows the concentrations of the insecticide mirex in 150 farmed salmon.

**Question:** Are the assumptions and conditions for inference satisfied?

- ✓ **Independence Assumption:** The fish were raised in many different places, and samples were purchased independently from several sources.
- ✓ **Randomization Condition:** The fish were selected randomly from those available for sale.



(continued)

<sup>7</sup> There are formal tests of Normality, but they don't really help. When we have a small sample—just when we really care about checking Normality—these tests have very little power. So it doesn't make much sense to use them in deciding whether to perform a  $t$ -test. We don't recommend that you use them.

For Example (continued)

- ✓ **10% Conditions:** There's lots of fish in the sea (and at the fish farms); 150 is certainly far fewer than 10% of the population.
- ✓ **Nearly Normal Condition:** The histogram of the data is unimodal. Although it may be somewhat skewed to the right, this is not a concern with a sample size of 150.

It's okay to use these data for inference about farm-raised salmon.



## JUST CHECKING

Every 10 years, the United States takes a census. The census tries to count every resident. There are two forms, known as the “short form,” answered by most people, and the “long form,” slogged through by about one in six or seven households chosen at random. According to the Census Bureau ([www.census.gov](http://www.census.gov)), “. . . each estimate based on the long form responses has an associated confidence interval.”

1. Why does the Census Bureau need a confidence interval for long-form information but not for the questions that appear on both the long and short forms?
2. Why must the Census Bureau base these confidence intervals on  $t$ -models?

The Census Bureau goes on to say, “These confidence intervals are wider . . . for geographic areas with smaller populations and for characteristics that occur less frequently in the area being examined (such as the proportion of people in poverty in a middle-income neighborhood).”

3. Why is this so? For example, why should a confidence interval for the mean amount families spend monthly on housing be wider for a sparsely populated area of farms in the Midwest than for a densely populated area of an urban center? How does the formula show this will happen?

To deal with this problem, the Census Bureau reports long-form data only for “. . . geographic areas from which about two hundred or more long forms were completed—which are large enough to produce good quality estimates. If smaller weighting areas had been used, the confidence intervals around the estimates would have been significantly wider, rendering many estimates less useful . . .”

4. Suppose the Census Bureau decided to report on areas from which only 50 long forms were completed. What effect would that have on a 95% confidence interval for, say, the mean cost of housing? Specifically, which values used in the formula for the margin of error would change? Which would change a lot and which would change only slightly?
5. Approximately how much wider would that confidence interval based on 50 forms be than the one based on 200 forms?

### STEP-BY-STEP EXAMPLE

### A One-Sample $t$ -Interval for the Mean

Let's build a 90% confidence interval for the mean speed of all vehicles traveling on Triphammer Road. The interval that we'll make is called the **one-sample  $t$ -interval**.

**Question:** What can we say about the mean speed of all cars on Triphammer Road?



**Plan** State what we want to know. Identify the parameter of interest.

Identify the variables and review the  $W$ 's.

I want to find a 90% confidence interval for the mean speed,  $\mu$ , of vehicles driving on Triphammer Road. I have data on the speeds of 23 cars there, sampled on April 11, 2000.

Make a picture. Check the distribution shape and look for skewness, multiple modes, and outliers.



The histogram centers around 30 mph, and the data lie between 20 and 40 mph. We'd expect a confidence interval to place the population mean within a few mph of 30.

**Model** Think about the assumptions and check the conditions.

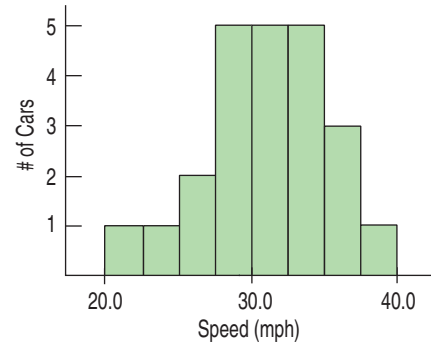
Note that with this small sample we probably didn't need to check the 10% Condition.

On the other hand, doing so gives us a chance to think about what the population is.

State the sampling distribution model for the statistic.

Choose your method.

Here's a histogram of the 23 observed speeds.



- ✓ **Independence Assumption:** This is a convenience sample, but care was taken to select cars that were not driving near each other, so their speeds are plausibly independent.
- ✓ **Randomization Condition:** Not really met. This is a convenience sample, but I have reason to believe that it is representative.
- ✓ **10% Condition:** The cars I observed were fewer than 10% of all cars that travel Triphammer Road.
- ✓ **Nearly Normal Condition:** The histogram of the speeds is unimodal and symmetric.

The conditions are satisfied, so I will use a Student's *t*-model with

$$(n - 1) = 22 \text{ degrees of freedom}$$

and find a **one-sample *t*-interval for the mean.**



**Mechanics** Construct the confidence interval.

Be sure to include the units along with the statistics.

The critical value we need to make a 90% interval comes from a Student's *t* table, a computer program, or a calculator. We have  $23 - 1 = 22$  degrees of freedom. The selected confidence level says that we want 90% of the probability to be caught in the middle, so we exclude 5% in *each* tail, for a total of 10%. The degrees

Calculating from the data (see page 530):

$$\begin{aligned} n &= 23 \text{ cars} \\ \bar{y} &= 31.0 \text{ mph} \\ s &= 4.25 \text{ mph.} \end{aligned}$$

The standard error of  $\bar{y}$  is

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{4.25}{\sqrt{23}} = 0.886 \text{ mph.}$$

The 90% critical value is  $t^*_{22} = 1.717$ , so the margin of error is

$$\begin{aligned} ME &= t^*_{22} \times SE(\bar{y}) \\ &= 1.717(0.886) \\ &= 1.521 \text{ mph.} \end{aligned}$$

The 90% confidence interval for the mean speed is  $31.0 \pm 1.5$  mph.

of freedom and 5% tail probability are all we need to know to find the critical value.



The result looks plausible and in line with what we thought.



**Conclusion** Interpret the confidence interval in the proper context.

When we construct confidence intervals in this way, we expect 90% of them to cover the true mean and 10% to miss the true value. That's what "90% confident" means.

I am 90% confident that the interval from 29.5 mph to 32.5 mph contains the true mean speed of all vehicles on Triphammer Road.

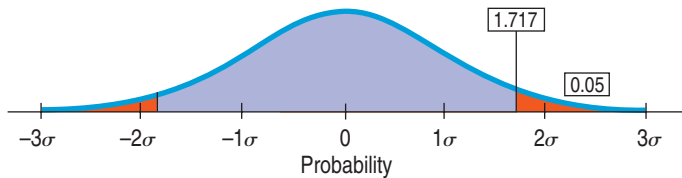
*Caveat:* This was not a random sample of vehicles. It was a convenience sample taken at one time on one day. And the participants were not blinded. Drivers could see the police device, and some may have slowed down. I'm reluctant to extend this inference to other situations.

**TI-*n*spire**

**Intervals for Means.** Generate confidence intervals from many samples to see how often they successfully capture the true mean.

Here's the part of the Student's *t* table that gives the critical value we needed for the Step-by-Step confidence interval. (See Table T in the back of the book.) To find a critical value, locate the row of the table corresponding to the degrees of freedom and the column corresponding to the probability you want. Our 90% confidence interval leaves 5% of the values on either side, so look for 0.05 at the top of the column or 90% at the bottom. The value in the table at that intersection is the critical value we need: 1.717.

As degrees of freedom increase, the shape of Student's *t*-models changes more gradually. Table T at the back of the book includes degrees of freedom between 100 and 1000 selected so that you can pin down the P-value for just about any df. If your df's aren't listed, take the cautious approach by using the next lower value or use technology.



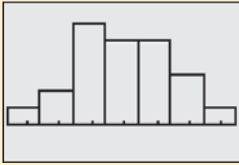
	0.25	0.2	0.15	0.1	0.05	0.025	0.02
19	.6876	.8610	1.066	1.328	1.729	2.093	2.205
20	.6870	.8600	1.064	1.325	1.725	2.086	2.197
21	.6864	.8591	1.063	1.323	1.721	2.080	2.189
22	.6858	.8583	1.061	1.321	1.717	2.074	2.183
23	.6853	.8575	1.060	1.319	1.714	2.069	2.177
24	.6848	.8569	1.059	1.318	1.711	2.064	2.172
25	.6844	.8562	1.058	1.316	1.708	2.060	2.167
26	.6840	.8557	1.058	1.315	1.706	2.056	2.162
27	.6837	.8551	1.057	1.314	1.703	2.052	2.158
C					80%	90%	95%

**A S** **Activity: Building *t*-Intervals with the *t*-Table.**

Interact with an animated version of Table T.

Of course, you can also create the confidence interval with computer software or a calculator.

## TI Tips



```

EDIT CALC TESTS
4:1-2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
8:TInterval...
9:2-SampZInt...
0:1-2-SampTInt...

```

```

TInterval
Inpt:Data Stats
List:L1
Freq:1
C-Level:.90
Calculate

```

```

TInterval
(29.523,32.564)
x=31.04347826
sx=4.247761559
n=23

```

```

TInterval
Inpt:Data Stats
x:83
sx:4
n:53
C-Level:.95
Calculate

```

```

TInterval
(81.897,84.103)
x=83
sx=4
n=53

```

## Finding a confidence interval for a mean

Yes, your calculator can create a confidence interval for a mean. And it's so easy we'll do two!

## Find a confidence interval given a set of data

- Type the speeds of the 23 Triphammer cars into **L1**. Go ahead; we'll wait.

```

29 34 34 28 30 29 38 31 29 34 32 31
27 37 29 26 24 34 36 31 34 36 21

```

- Set up a **STATPLOT** to create a histogram of the data so you can check the nearly Normal condition. Looks okay—unimodal and roughly symmetric.
- Under **STAT TESTS** choose **8:TInterval**.
- Choose **Inpt:Data**, then specify that your data is **List:L1**.
- For these data the frequency is 1. (If your data have a frequency distribution stored in another list, you would specify that.)
- Choose the confidence level you want.
- Calculate** the interval.

There's the 90% confidence interval. That was easy—but remember, the calculator only does the *Show*. Now you have to *Tell* what it means.

## No data? Find a confidence interval given the sample's mean and standard deviation

Sometimes instead of the original data you just have the summary statistics. For instance, suppose a random sample of 53 lengths of fishing line had a mean strength of 83 pounds and standard deviation of 4 pounds. Let's make a 95% confidence interval for the mean strength of this kind of fishing line.

- Without the data you can't check the Nearly Normal Condition. But 53 is a moderately large sample, so assuming there were no outliers, it's okay to proceed. You need to say that.
- Go back to **STAT TESTS** and choose **8:TInterval** again. This time indicate that you wish to enter the summary statistics. To do that, select **Stats**, then hit **ENTER**.
- Specify the sample mean, standard deviation, and sample size.
- Choose a confidence level and **Calculate** the interval.
- If (repeat, IF . . .) strengths of fishing lines follow a Normal model, we are 95% confident that this kind of line has a mean strength between 81.9 and 84.1 pounds.

## More Cautions About Interpreting Confidence Intervals

AS

**Activity: Intuition for  $t$ -based Intervals.** A narrated review of Student's  $t$ .

Confidence intervals for means offer new tempting wrong interpretations. Here are some things you *shouldn't* say:

- Don't say**, "90% of all the vehicles on Triphammer Road drive at a speed between 29.5 and 32.5 mph." The confidence interval is about the *mean* speed, not about the speeds of *individual* vehicles.

**So What Should We Say?**

Since 90% of random samples yield an interval that captures the true mean, we *should* say, “I am 90% confident that the interval from 29.5 to 32.5 mph contains the mean speed of all the vehicles on Triphammer Road.” It’s also okay to say something less formal: “I am 90% confident that the average speed of all vehicles on Triphammer Road is between 29.5 and 32.5 mph.” Remember: *Our uncertainty is about the interval, not the true mean.* The interval varies randomly. The true mean speed is neither variable nor random—just unknown.

- ▶ *Don’t say*, “We are 90% confident that a randomly selected vehicle will have a speed between 29.5 and 32.5 mph.” This false interpretation is also about individual vehicles rather than about the *mean* of the speeds. We are 90% confident that the *mean* speed of all vehicles on Triphammer Road is between 29.5 and 32.5 mph.
- ▶ *Don’t say*, “The mean speed of the vehicles is 31.0 mph 90% of the time.” That’s about means, but still wrong. It implies that the true mean varies, when in fact it is the confidence interval that would have been different had we gotten a different sample.
- ▶ Finally, *don’t say*, “90% of all samples will have mean speeds between 29.5 and 32.5 mph.” That statement suggests that *this* interval somehow sets a standard for every other interval. In fact, this interval is no more (or less) likely to be correct than any other. You could say that 90% of all possible samples will produce intervals that actually do contain the true mean speed. (The problem is that, because we’ll never know where the true mean speed really is, we can’t know if our sample was one of those 90%.)
- ▶ *Do say*, “90% of intervals that could be found in this way would cover the true value.” Or make it more personal and say, “I am 90% confident that the true mean speed is between 29.5 and 32.5 mph.”

## Make a Picture, Make a Picture, Make a Picture

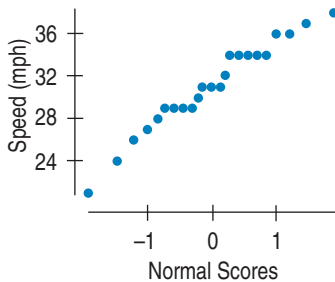


FIGURE 23.3

A Normal probability plot of speeds looks reasonably straight.

The only reasonable way to check the Nearly Normal Condition is with graphs of the data. Make a histogram of the data and verify that its distribution is unimodal and symmetric and that it has no outliers. You may also want to make a Normal probability plot to see that it’s reasonably straight. You’ll be able to spot deviations from the Normal model more easily with a Normal probability plot, but it’s easier to understand the particular nature of the deviations from a histogram.

If you have a computer or graphing calculator doing the work, there’s no excuse not to look at *both* displays as part of checking the Nearly Normal Condition.

## A Test for the Mean

The residents along Triphammer Road have a more specific concern. It appears that the mean speed along the road is higher than it ought to be. To get the police to patrol more frequently, though, they’ll need to show that the true mean speed is *in fact greater* than the 30 mph speed limit. This calls for a hypothesis test called the **one-sample *t*-test for the mean**.

You already know enough to construct this test. The test statistic looks just like the others we’ve seen. It compares the difference between the observed statistic and a hypothesized value to the standard error of the observed statistic. We already know that, for means, the appropriate probability model to use for P-values is Student’s *t* with  $n - 1$  degrees of freedom.

We're ready to go:

**AS** **Activity: A  $t$ -Test for Wind Speed.** Watch the video in the preceding activity, and then use the interactive tool to test whether there's enough wind for electricity generation at a site under investigation.

### ONE-SAMPLE $t$ -TEST FOR THE MEAN

The assumptions and conditions for the one-sample  $t$ -test for the mean are the same as for the one-sample  $t$ -interval. We test the hypothesis  $H_0: \mu = \mu_0$  using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}.$$

The standard error of  $\bar{y}$  is  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ .

When the conditions are met and the null hypothesis is true, this statistic follows a Student's  $t$ -model with  $n - 1$  degrees of freedom. We use that model to obtain a P-value.

### FOR EXAMPLE

#### A one-sample $t$ -test for the mean

**Recap:** Researchers tested 150 farm-raised salmon for organic contaminants. They found the mean concentration of the carcinogenic insecticide mirex to be 0.0913 parts per million, with standard deviation 0.0495 ppm. As a safety recommendation to recreational fishers, the Environmental Protection Agency's (EPA) recommended "screening value" for mirex is 0.08 ppm.

**Question:** Are farmed salmon contaminated beyond the level permitted by the EPA? (We've already checked the conditions; see pages 537–8.)

$$H_0: \mu = 0.08$$

$$H_A: \mu > 0.08$$

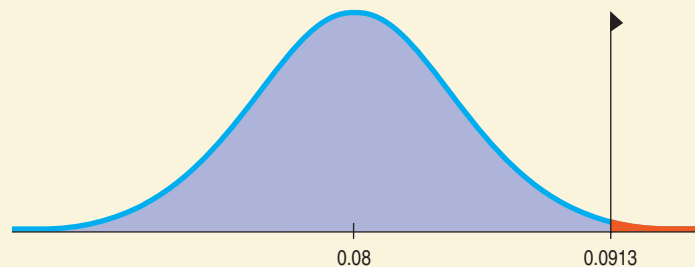
These data satisfy the conditions for inference; I'll do a one-sample  $t$ -test for the mean:

$$n = 150, df = 149$$

$$\bar{y} = 0.0913, s = 0.0495$$

$$SE(\bar{y}) = \frac{0.0495}{\sqrt{150}} = 0.0040$$

$$t_{149} = \frac{0.0913 - 0.08}{0.0040} = 2.825$$



$$P(t_{149} > 2.825) = 0.0027 \text{ (from technology).}$$

With a P-value that low, I reject the null hypothesis and conclude that, in farm-raised salmon, the mirex contamination level does exceed the EPA screening value.

### STEP-BY-STEP EXAMPLE

#### A One-Sample $t$ -Test for the Mean

Let's apply the one-sample  $t$ -test to the Triphammer Road car speeds. The speed limit is 30 mph, so we'll use that as the null hypothesis value.

**Question:** Does the mean speed of all cars exceed the posted speed limit?

**THINK**

**Plan** State what we want to know. Make clear what the population and parameter are.

Identify the variables and review the W's.

**Hypotheses** The null hypothesis is that the true mean speed is equal to the limit. Because we're interested in whether the vehicles are speeding, the alternative is one-sided.

Make a picture. Check the distribution for skewness, multiple modes, and outliers.

**REALITY CHECK**

The histogram of the observed speeds is clustered around 30, so we'd be surprised to find that the mean was much higher than that. (The fact that 30 is within the confidence interval that we've just found confirms this suspicion.)

**Model** Think about the assumptions and check the conditions.

(We won't worry about the 10% Condition—it's a small sample.)

State the sampling distribution model. (Be sure to include the degrees of freedom.)

Choose your method.

**SHOW**

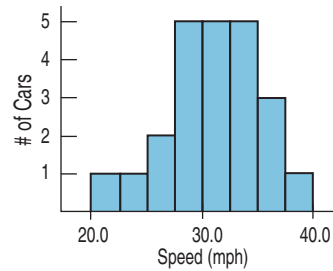
**Mechanics** Be sure to include the units when you write down what you know from the data.

We use the null model to find the P-value. Make a picture of the  $t$ -model centered at  $\mu = 30$ . Since this is an upper-tail test, shade the region to the right of the observed mean speed.

I want to know whether the mean speed of vehicles on Triphammer Road exceeds the posted speed limit of 30 mph. I have a sample of 23 car speeds on April 11, 2000.

$$H_0: \text{Mean speed, } \mu = 30 \text{ mph}$$

$$H_A: \text{Mean speed, } \mu > 30 \text{ mph}$$



- ✓ **Independence Assumption:** These cars are a convenience sample, but they were selected so no two cars were driving near each other, so I am justified in believing that their speeds are independent.
- ✓ **Randomization Condition:** Although I have a convenience sample, I have reason to believe that it is a representative sample.
- ✓ **Nearly Normal Condition:** The histogram of the speeds is unimodal and reasonably symmetric.

The conditions are satisfied, so I'll use a Student's  $t$ -model with  $(n - 1) = 22$  degrees of freedom to do a **one-sample  $t$ -test for the mean**.

From the data,

$$n = 23 \text{ cars}$$

$$\bar{y} = 31.0 \text{ mph}$$

$$s = 4.25 \text{ mph}$$

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} = \frac{4.25}{\sqrt{23}} = 0.886 \text{ mph.}$$

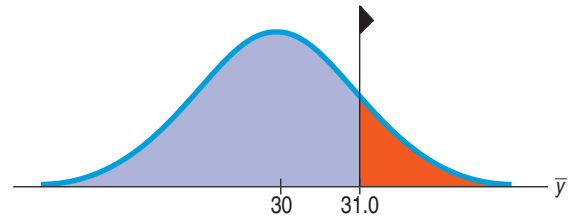


The  $t$ -statistic calculation is just a standardized value, like  $z$ . We subtract the hypothesized mean and divide by the standard error.

The P-value is the probability of observing a sample mean as large as 31.0 (or larger) if the true mean were 30.0, as the null hypothesis states. We can find this P-value from a table, calculator, or computer program.

**REALITY CHECK**

We're not surprised that the difference isn't statistically significant.



$$t = \frac{\bar{y} - \mu_0}{SE(\bar{y})} = \frac{31.0 - 30.0}{0.886} = 1.13$$

(The observed mean is 1.13 standard errors above the hypothesized value.)

$$P\text{-value} = P(t_{22} > 1.13) = 0.136$$



**Conclusion** Link the P-value to your decision about  $H_0$ , and state your conclusion in context.

Unfortunately for the residents, there is no course of action associated with failing to reject this particular null hypothesis.

The P-value of 0.136 says that if the true mean speed of vehicles on Triphammer Road were 30 mph, samples of 23 vehicles can be expected to have an observed mean of at least 31.0 mph 13.6% of the time. That P-value is not small enough for me to reject the hypothesis that the true mean is 30 mph at any reasonable alpha level. I conclude that there is not enough evidence to say the average speed is too high.

**TI Tips**

**Testing a hypothesis about a mean**

```
EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:Interval...
```

```
T-Test
Inpt:DATA Stats
μ₀:30
List:L₁
Freq:1
μ:≠μ₀ <μ₀ >μ₀
Calculate Draw
```

```
T-Test
μ>30
t=1.178113665
P=.1256691057
x̄=31.04347826
Sx=4.247761559
n=23
```

**Testing a Hypothesis Given a Set of Data**

Still have the Triphammer Road auto speeds in **L1**? Good. Let's use the TI to see if the mean is significantly higher than 30 mph (you've already checked the histogram to verify the nearly Normal condition, of course).

- Go to the **STAT TESTS** menu, and choose **2:T-Test**.
- Tell it you want to use the stored **Data**.
- Enter the mean of the null model, and indicate where the data are.
- Since this is an upper tail test, choose the  $\mu > \mu_0$  option.
- **Calculate**.

There's everything you need to know: the summary statistics, the calculated value of  $t$ , and the P-value of 0.126. ( $t$  and  $P$  differ slightly from the values in our worked example because when we did it by hand we rounded off the mean and standard deviation. No harm done.)

As always, the *Tell* is up to you.

```
T-Test
Inpt:Data Stats
μ₀:80
x̄:83
sₓ:4
n:53
μ:≠μ₀ <μ₀ >μ₀
Calculate Draw
```

```
T-Test
μ>80
t=5.460082417
P=6.7566262E-7
x̄=83
sₓ=4
n=53
```

### Testing a Hypothesis Given the Sample's Mean and Standard Deviation

Don't have the actual data? Just summary statistics? No problem, assuming you can verify the necessary conditions. In the last TI Tips we created a confidence interval for the strength of fishing line. We had test results for a random sample of 53 lengths of line showing a mean strength of 83 pounds and a standard deviation of 4 pounds. Is there evidence that this kind of fishing line exceeds the "80-lb test" as labeled on the package?

We bet you know what to do even without our help. Try it before you read on.

- Go back to **2:T-Test**.
- You're entering **Stats** this time.
- Specify the hypothesized mean and the sample statistics.
- Choose the alternative being tested (upper tail here).
- **Calculate**.

The results of the calculator's mechanics show a large  $t$  and a really small  $P$ -value (0.0000007). We have very strong evidence that the mean breaking strength of this kind of fishing line is over the 80 pounds claimed by the manufacturer.

## Significance and Importance

Recall that "statistically significant" does not mean "actually important" or "meaningful," even though it sort of sounds that way. In this example, it does seem that speeds may be a bit above 30 miles per hour. If so, it's possible that a larger sample would show statistical significance.

But would that be the right decision? The difference between 31 miles per hour and 30 miles per hour doesn't seem meaningful, and rejecting the null hypothesis wouldn't change that. Even with a statistically significant result, it would be hard to convince the police that vehicles on Triphammer Road were driving at dangerously fast speeds. It would probably also be difficult to persuade the town that spending more money to lower the average speed on Triphammer Road would be a good use of the town's resources. Looking at the confidence interval, we can say with 90% confidence that the mean speed is somewhere between 29.5 and 32.5 mph. Even in the worst case, if the mean speed is 32.5 mph, would this be a bad enough situation to convince the town to spend more money? Probably not. It's always a good idea when we test a hypothesis to also check the confidence interval and think about the likely values for the mean.



### JUST CHECKING

In discussing estimates based on the long-form samples, the Census Bureau notes, "The disadvantage . . . is that . . . estimates of characteristics that are also reported on the short form will not match the [long-form estimates]."

The short-form estimates are values from a complete census, so they are the "true" values—something we don't usually have when we do inference.

6. Suppose we use long-form data to make 95% confidence intervals for the mean age of residents for each of 100 of the Census-defined areas. How many of these 100 intervals should we expect will fail to include the true mean age (as determined from the complete short-form Census data)?
7. Based only on the long-form sample, we might test the null hypothesis about the mean household income in a region. Would the power of the test increase or decrease if we used an area with more long forms?

## Intervals and Tests

The 90% confidence interval for the mean speed was  $31.0 \text{ mph} \pm 1.5$ , or (29.5 mph, 32.5 mph). If someone hypothesized that the mean speed was really 30 mph, how would you feel about it? How about 35 mph?

Because the confidence interval included the speed limit of 30 mph, it certainly looked like 30 mph might be a plausible value for the true mean speed of the vehicles on Triphammer Road. In fact, 30 mph gave a P-value of 0.136—too large to reject the null hypothesis. We should have seen this coming. The hypothesized mean of 30 mph lies *within the confidence interval*. It's one of the reasonable values for the mean.

Confidence intervals and significance tests are built from the same calculations. In fact, they are really complementary ways of looking at the same question. Here's the connection: The confidence interval contains all the null hypothesis values we can't reject with these data.

More precisely, a level  $C$  confidence interval contains *all* of the plausible null hypothesis values that would *not* be rejected by a two-sided hypothesis test at alpha level  $1 - C$ . So a 95% confidence interval matches a  $1 - 0.95 = 0.05$  level two-sided test for these data.

Confidence intervals are naturally two-sided, so they match exactly with two-sided hypothesis tests. When, as in our example, the hypothesis is one-sided, the corresponding alpha level is  $(1 - C)/2$ .

**Fail to reject** Our 90% confidence interval was 29.5 to 32.5 mph. If any of these values had been the null hypothesis for the mean, then the corresponding hypothesis test at  $\alpha = 0.05$  (because  $\frac{1 - 0.90}{2} = 0.05$ ) would not have been able to reject the null. That is, the corresponding one-sided P-value for our observed mean of 31 mph would be greater than 0.05. So, we would not reject any hypothesized value between 29.5 and 32.5 mph.

## Sample Size

**AS**

**Activity: The Real Effect of Small Sample Size.** We know that smaller sample sizes lead to wider confidence intervals, but is that just because they have fewer degrees of freedom?

How large a sample do we need? The simple answer is “more.” But more data cost money, effort, and time, so how much is enough? Suppose your computer just took an hour to download a movie you wanted to watch. You're not happy. You hear about a program that claims to download movies in under a half hour. You're interested enough to spend \$29.95 for it, but only if it really delivers. So you get the free evaluation copy and test it by downloading that movie 5 different times. Of course, the mean download time is not exactly 30 minutes as claimed. Observations vary. If the margin of error were 8 minutes, though, you'd probably be able to decide whether the software is worth the money. Doubling the sample size would require another 5 hours of testing and would reduce your margin of error to a bit under 6 minutes. You'll need to decide whether that's worth the effort.

As we make plans to collect data, we should have some idea of how small a margin of error we need to be able to draw a conclusion or detect a difference we want to see. If the size of the effect we're studying is large, then we may be able to tolerate a larger *ME*. If we need great precision, however, we'll want a smaller *ME*, and, of course, that means a larger sample size.

Armed with the *ME* and confidence level, we can find the sample size we'll need. Almost.

We know that for a mean,  $ME = t_{n-1}^* \times SE(\bar{y})$  and that  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ , so we can determine the sample size by solving this equation for  $n$ :

$$ME = t_{n-1}^* \frac{s}{\sqrt{n}}$$

The good news is that we have an equation; the bad news is that we won't know most of the values we need to solve it. When we thought about sample size for proportions back in Chapter 19, we ran into a similar problem. There we had to guess a working value for  $p$  to compute a sample size. Here, we need to know  $s$ . We don't know  $s$  until we get some data, but we want to calculate the sample size *before* collecting the data. We might be able to make a good guess, and that is often good enough for this purpose. If we have no idea what the standard deviation might be, or if the sample size really matters (for example, because each additional individual is very expensive to sample or experiment on), it might be a good idea to run a small *pilot study* to get some feeling for the standard deviation.

That's not all. Without knowing  $n$ , we don't know the degrees of freedom and we can't find the critical value,  $t_{n-1}^*$ . One common approach is to use the corresponding  $z^*$  value from the Normal model. If you've chosen a 95% confidence level, then just use 2, following the 68–95–99.7 Rule. If your estimated sample size is, say, 60 or more, it's probably okay— $z^*$  was a good guess. If it's smaller than that, you may want to add a step, using  $z^*$  at first, finding  $n$ , and then replacing  $z^*$  with the corresponding  $t_{n-1}^*$  and calculating the sample size once more.

Sample size calculations are *never* exact. The margin of error you find *after* collecting the data won't match exactly the one you used to find  $n$ . The sample size formula depends on quantities that you won't have until you collect the data, but using it is an important first step. Before you collect data, it's always a good idea to know whether the sample size is large enough to give you a good chance of being able to tell you what you want to know.

## FOR EXAMPLE

### Finding sample size

A company claims its program will allow your computer to download movies quickly. We'll test the free evaluation copy by downloading a movie several times, hoping to estimate the mean download time with a margin of error of only 8 minutes. We think the standard deviation of download times is about 10 minutes.

**Question:** How many trial downloads must we run if we want 95% confidence in our estimate with a margin of error of only 8 minutes?

Using  $z^* = 1.96$ , solve

$$\begin{aligned} 8 &= 1.96 \frac{10}{\sqrt{n}} \\ \sqrt{n} &= \frac{1.96 \times 10}{8} = 2.45 \\ n &= (2.45)^2 = 6.0025 \end{aligned}$$

That's a small sample size, so I'll use  $(6 - 1) = 5$  degrees of freedom<sup>8</sup> to substitute an appropriate  $t^*$  value. At 95%,  $t_5^* = 2.571$ . Solving the equation one more time:

$$8 = 2.571 \frac{10}{\sqrt{n}}$$

<sup>8</sup> Ordinarily we'd round the sample size *up*. But at this stage of the calculation, rounding *down* is the safer choice. Can you see why?

$$\sqrt{n} = \frac{2.571 \times 10}{8} \approx 3.214$$

$$n = (3.214)^2 \approx 10.33$$

To make sure the ME is no larger, I'll round up, which gives  $n = 11$  runs. So, to get an ME of 8 minutes, I'll find the downloading times for 11 movies.

## Degrees of Freedom

Some calculators offer an alternative button for standard deviation that divides by  $n$  instead of  $n - 1$ . Why don't you stick a wad of gum over the "n" button so you won't be tempted to use it? Use  $n - 1$ .

The number of degrees of freedom,  $(n - 1)$ , might have reminded you of the value we divide by to find the standard deviation of the data (since, in fact, it's the same number). When we introduced that formula, we promised to say a bit more about why we divide by  $n - 1$  rather than by  $n$ . The reason is closely tied to the reasoning behind the  $t$ -distribution.

If only we knew the true population mean,  $\mu$ , we would find the sample standard deviation as

$$s = \sqrt{\frac{\sum (y - \mu)^2}{n}} \quad (\text{Equation 23.1})^9$$

We use  $\bar{y}$  instead of  $\mu$ , though, and that causes a problem. For any sample, the data values will generally be closer to their own sample mean than to the true population mean,  $\mu$ . Why is that? Imagine that we take a random sample of 10 high school seniors. The mean SAT verbal score is 500 in the United States. But the sample mean,  $\bar{y}$ , for *these* 10 seniors won't be exactly 500. Are the 10 seniors' scores closer to 500 or  $\bar{y}$ ? They'll always be closer to their own average  $\bar{y}$ . If we used  $\sum (y - \bar{y})^2$  instead of  $\sum (y - \mu)^2$  in Equation 23.1 to calculate  $s$ , our standard deviation estimate would be too small. How can we fix it? The amazing mathematical fact is that we can compensate for the smaller sum exactly by dividing by  $n - 1$  instead of by  $n$ . So that's all the  $n - 1$  is doing in the denominator of  $s$ . And we call  $n - 1$  the degrees of freedom.

### WHAT CAN GO WRONG?

The most fundamental issue you face is knowing when to use Student's  $t$  methods.

- ▶ **Don't confuse proportions and means.** When you treat your data as categorical, counting successes and summarizing with a sample proportion, make inferences using the Normal model methods you learned about in Chapters 19 through 22. When you treat your data as quantitative, summarizing with a sample mean, make your inferences using Student's  $t$  methods.

Student's  $t$  methods work only when the Normality Assumption is true. Naturally, many of the ways things can go wrong turn out to be different ways that the Normality

(continued)

<sup>9</sup> Statistics textbooks usually have equation numbers so they can talk about equations by name. We haven't needed equation numbers yet, but we admit it's useful here, so this is our first.

As tempting as it is to get rid of annoying values, you can't just throw away outliers and not discuss them. It isn't appropriate to lop off the highest or lowest values just to improve your results.

Assumption can fail. It's always a good idea to look for the most common kinds of failure. It turns out that you can even fix some of them.

- ▶ **Beware of multimodality.** The Nearly Normal Condition clearly fails if a histogram of the data has two or more modes. When you see this, look for the possibility that your data come from two groups. If so, your best bet is to try to separate the data into different groups. (Use the variables to help distinguish the modes, if possible. For example, if the modes seem to be composed mostly of men in one and women in the other, split the data according to sex.) Then you could analyze each group separately.
- ▶ **Beware of skewed data.** Make a Normal probability plot and a histogram of the data. If the data are very skewed, you might try re-expressing the variable. Re-expressing may yield a distribution that is unimodal and symmetric, more appropriate for Student's  $t$  inference methods for means. Re-expression cannot help if the sample distribution is not unimodal. Some people may object to re-expressing the data, but unless your sample is very large, you just can't use the methods of this chapter on skewed data.
- ▶ **Set outliers aside.** Student's  $t$  methods are built on the mean and standard deviation, so we should beware of outliers when using them. When you make a histogram to check the Nearly Normal Condition, be sure to check for outliers as well. If you find some, consider doing the analysis twice, both with the outliers excluded and with them included in the data, to get a sense of how much they affect the results.

The suggestion that you can perform an analysis with outliers removed may be controversial in some disciplines. Setting aside outliers is seen by some as "cheating." But an analysis of data with outliers left in place is *always* wrong. The outliers violate the Nearly Normal Condition and also the implicit assumption of a homogeneous population, so they invalidate inference procedures. An analysis of the non-outlying points, along with a separate discussion of the outliers, is often much more informative and can reveal important aspects of the data.

How can you tell whether there are outliers in your data? The "outlier nomination rule" of boxplots can offer some guidance, but it's just a rule of thumb and not an absolute definition. The best practical definition is that a value is an outlier if removing it substantially changes your conclusions about the data. You won't want a single value to determine your understanding of the world unless you are very, very sure that it is absolutely correct. Of course, when the outliers affect your conclusion, this can lead to the uncomfortable state of not really knowing what to conclude. Such situations call for you to use your knowledge of the real world and your understanding of the data you are working with.<sup>10</sup>

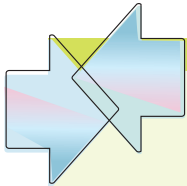
Of course, Normality issues aren't the only risks you face when doing inferences about means. Remember to *Think* about the usual suspects.

- ▶ **Watch out for bias.** Measurements of all kinds can be biased. If your observations differ from the true mean in a systematic way, your confidence interval may not capture the true mean. And there is no sample size that will save you. A bathroom scale that's 5 pounds off will be 5 pounds off even if you weigh yourself 100 times and take the average. We've seen several sources of bias in surveys, and measurements can be biased, too. Be sure to think about possible sources of bias in your measurements.
- ▶ **Make sure cases are independent.** Student's  $t$  methods also require the sampled values to be mutually independent. We check for random sampling and the 10% Condition. You should also think hard about whether there are likely violations of independence in the data collection method. If there are, be very cautious about using these methods.
- ▶ **Make sure that data are from an appropriately randomized sample.** Ideally, all data that we analyze are drawn from a simple random sample or generated by a randomized experiment. When they're not, be careful about making inferences from them. You

<sup>10</sup> An important reason for you to know Statistics rather than let someone else analyze your data.

may still compute a confidence interval correctly, or get the mechanics of the P-value right, but this might not save you from making a serious mistake in inference.

- ▶ **Interpret your confidence interval correctly.** Many statements that sound tempting are, in fact, misinterpretations of a confidence interval for a mean. You might want to have another look at some of the common mistakes, explained on pages 541–2. Keep in mind that a confidence interval is about the mean of the population, not about the means of samples, individuals in samples, or individuals in the population.



## CONNECTIONS

The steps for finding a confidence interval or hypothesis test for means are just like the corresponding steps for proportions. Even the form of the calculations is similar. As the  $z$ -statistic did for proportions, the  $t$ -statistic tells us how many standard errors our sample mean is from the hypothesized mean. For means, though, we have to estimate the standard error separately. This added uncertainty changes the model for the sampling distribution from  $z$  to  $t$ .

As with all of our inference methods, the randomization applied in drawing a random sample or in randomizing a comparative experiment is what generates the sampling distribution. Randomization is what makes inference in this way possible at all.

The new concept of degrees of freedom connects back to the denominator of the sample standard deviation calculation, as shown earlier.

There's just no escaping histograms and Normal probability plots. The Nearly Normal Condition required to use Student's  $t$  can be checked best by making appropriate displays of the data. Back when we first used histograms, we looked at their shape and, in particular, checked whether they were unimodal and symmetric, and whether they showed any outliers. Those are just the features we check for here. The Normal probability plot zeros in on the Normal model a little more precisely.

## WHAT HAVE WE LEARNED?



We first learned to create confidence intervals and test hypotheses about proportions. Now we've turned our attention to means, and learned that statistical inference for means relies on the same concepts; only the mechanics and our model have changed.

- ▶ We've learned that what we can say about a population mean is inferred from data, using the mean of a representative random sample.
- ▶ We've learned to describe the sampling distribution of sample means using a new model we select from the Student's  $t$  family based on our degrees of freedom.
- ▶ We've learned that our ruler for measuring the variability in sample means is the standard error  $SE(\bar{y}) = \frac{s}{\sqrt{n}}$ .
- ▶ We've learned to find the margin of error for a confidence interval using that ruler and critical values based on a Student's  $t$ -model.
- ▶ And we've also learned to use that ruler to test hypotheses about the population mean.

Above all, we've learned that the reasoning of inference, the need to verify that the appropriate assumptions are met, and the proper interpretation of confidence intervals and P-values all remain the same regardless of whether we are investigating means or proportions.

## Terms

Student's  $t$   
Degrees of freedom (df)

533. A family of distributions indexed by its degrees of freedom. The  $t$ -models are unimodal symmetric, and bell shaped, but generally have fatter tails and a narrower center than the Normal model. As the degrees of freedom increase,  $t$ -distributions approach the Normal.

One-sample  $t$ -interval  
for the mean

534. A one-sample  $t$ -interval for the population mean is

$$\bar{y} \pm t_{n-1}^* \times SE(\bar{y}), \text{ where } SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

The critical value  $t_{n-1}^*$  depends on the particular confidence level,  $C$ , that you specify and on the number of degrees of freedom,  $n - 1$ .

One-sample  $t$ -test for  
the mean

543. The one-sample  $t$ -test for the mean tests the hypothesis  $H_0: \mu = \mu_0$  using the statistic

$$t_{n-1} = \frac{\bar{y} - \mu_0}{SE(\bar{y})}.$$

The standard error of  $\bar{y}$  is

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}.$$

## Skills

THINK

- ▶ Know the assumptions required for  $t$ -tests and  $t$ -based confidence intervals.
- ▶ Know how to examine your data for violations of conditions that would make inference about the population mean unwise or invalid.
- ▶ Understand that a confidence interval and a hypothesis test are essentially equivalent. You can do a two-tailed hypothesis test at level of significance  $\alpha$  with a  $1 - \alpha$  confidence interval, or a one-tailed test with a  $1 - 2\alpha$  confidence interval.

SHOW

- ▶ Be able to compute and interpret a  $t$ -test for the population mean using a statistics package or working from summary statistics for a sample.
- ▶ Be able to compute and interpret a  $t$ -based confidence interval for the population mean using a statistics package or working from summary statistics for a sample.

TELL

- ▶ Be able to explain the meaning of a confidence interval for a population mean. Make clear that the randomness associated with the confidence level is a statement about the interval bounds and not about the population parameter value.
- ▶ Understand that a 95% confidence interval does not trap 95% of the sample values.
- ▶ Be able to interpret the result of a test of a hypothesis about a population mean.
- ▶ Know that we do not “accept” a null hypothesis if we cannot reject it. We say that we fail to reject it.
- ▶ Understand that the P-value of a test does not give the probability that the null hypothesis is correct.

## INFERENCE FOR MEANS ON THE COMPUTER

Statistics packages offer convenient ways to make histograms of the data. Even better for assessing near-Normality is a Normal probability plot. When you work on a computer, there is simply no excuse for skipping the step of plotting the data to check that it is nearly Normal. Beware: Statistics packages don't agree on whether to place the Normal scores on the x-axis (as we have done) or the y-axis. Read the axis labels.



Any standard statistics package can compute a hypothesis test. Here's what the package output might look like in general (although no package we know gives the results in exactly this form):<sup>11</sup>

**AS** **Activity: Student's *t* in Practice.** We almost always use technology to do inference with Student's *t*. Here's a chance to do that as you investigate several questions.

Null hypothesis Alternative hypothesis

Test Ho:  $\mu$  (speed) = 30 vs Ha:  $\mu$  (speed) > 30  
 Sample Mean = 31.043478  
 $t = 1.178$  w/22 df  
 P-value = 0.1257

The *t*-statistic (and its degrees of freedom)

The P-value is usually given last

The package computes the sample mean and sample standard deviation of the variable and finds the P-value from the *t*-distribution based on the appropriate number of degrees of freedom. All modern statistics packages report P-values. The package may also provide additional information such as the sample mean, sample standard deviation, *t*-statistic value, and degrees of freedom. These are useful for interpreting the resulting P-value and telling the difference between a meaningful result and one that is merely statistically significant. Statistics packages that report the estimated standard deviation of the sampling distribution usually label it "standard error" or "SE." Inference results are also sometimes reported in a table. You may have to read carefully to find the values you need. Often, test results and the corresponding confidence interval bounds are given together. And often you must read carefully to find the alternative hypotheses. Here's an example of that kind of output:

$\mu_0$  Calculated mean,  $\bar{y}$

Hypothesized value	30		
Estimated mean	31.043478		
DF	22		
Std Error	0.886		
Alpha	0.05		

	tTest	t interval	
Statistic	1.178		
Prob >  t	0.2513	Upper 95%	32.880348
Prob > t	0.1257	Lower 95%	29.206608
Prob < t	0.8743		

The alpha level often defaults to 0.05. Some packages let you choose a different alpha level

t-statistic

P-values for each alternative

Corresponding confidence interval

2-sided alternative (note the |t|)

1-sided  $H_A: \mu > 30$

1-sided  $H_A: \mu < 30$

The commands to do inference for means on common statistics programs and calculators are not always obvious. (By contrast, the resulting output is usually clearly labeled and easy to read.) The guides for each program can help you start navigating.

<sup>11</sup> Many statistics packages keep as many as 16 digits for all intermediate calculations. If we had kept as many, our results in the Step-By-Step section would have been closer to these.

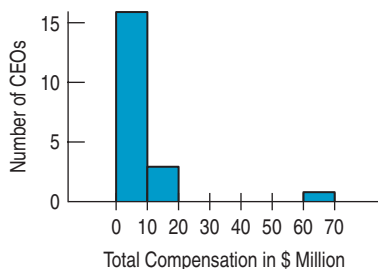
## EXERCISES

1. ***t*-models, part I.** Using the *t* tables, software, or a calculator, estimate
- the critical value of *t* for a 90% confidence interval with *df* = 17.
  - the critical value of *t* for a 98% confidence interval with *df* = 88.
  - the P-value for  $t \geq 2.09$  with 4 degrees of freedom.
  - the P-value for  $|t| > 1.78$  with 22 degrees of freedom.
2. ***t*-models, part II.** Using the *t* tables, software, or a calculator, estimate
- the critical value of *t* for a 95% confidence interval with *df* = 7.
  - the critical value of *t* for a 99% confidence interval with *df* = 102.
  - the P-value for  $t \leq 2.19$  with 41 degrees of freedom.
  - the P-value for  $|t| > 2.33$  with 12 degrees of freedom.
3. ***t*-models, part III.** Describe how the shape, center, and spread of *t*-models change as the number of degrees of freedom increases.
4. ***t*-models, part IV (last one!).** Describe how the critical value of *t* for a 95% confidence interval changes as the number of degrees of freedom increases.
5. **Cattle.** Livestock are given a special feed supplement to see if it will promote weight gain. Researchers report that the 77 cows studied gained an average of 56 pounds, and that a 95% confidence interval for the mean weight gain this supplement produces has a margin of error of  $\pm 11$  pounds. Some students wrote the following conclusions. Did anyone interpret the interval correctly? Explain any misinterpretations.
- 95% of the cows studied gained between 45 and 67 pounds.
  - We're 95% sure that a cow fed this supplement will gain between 45 and 67 pounds.
  - We're 95% sure that the average weight gain among the cows in this study was between 45 and 67 pounds.
  - The average weight gain of cows fed this supplement will be between 45 and 67 pounds 95% of the time.
  - If this supplement is tested on another sample of cows, there is a 95% chance that their average weight gain will be between 45 and 67 pounds.
6. **Teachers.** Software analysis of the salaries of a random sample of 288 Nevada teachers produced the confidence interval shown below. Which conclusion is correct? What's wrong with the others?
- |                                                                                                          |                                                                               |
|----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| $t\text{-Interval for } \mu: \quad \text{with 90.00\% Confidence,}$ $38944 < \mu(\text{TchPay}) < 42893$ | $\text{With 95.00\% Confidence,}$ $70.887604 < \mu(\text{Pulse}) < 74.497011$ |
|----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|
- If we took many random samples of 288 Nevada teachers, about 9 out of 10 of them would produce this confidence interval.
  - If we took many random samples of Nevada teachers, about 9 out of 10 of them would produce a confidence interval that contained the mean salary of all Nevada teachers.
  - About 9 out of 10 Nevada teachers earn between \$38,944 and \$42,893.
  - About 9 out of 10 of the teachers surveyed earn between \$38,944 and \$42,893.
  - We are 90% confident that the average teacher salary in the United States is between \$38,944 and \$42,893.
7. **Meal plan.** After surveying students at Dartmouth College, a campus organization calculated that a 95% confidence interval for the mean cost of food for one term (of three in the Dartmouth trimester calendar) is (\$1102, \$1290). Now the organization is trying to write its report and is considering the following interpretations. Comment on each.
- 95% of all students pay between \$1102 and \$1290 for food.
  - 95% of the sampled students paid between \$1102 and \$1290.
  - We're 95% sure that students in this sample averaged between \$1102 and \$1290 for food.
  - 95% of all samples of students will have average food costs between \$1102 and \$1290.
  - We're 95% sure that the average amount all students pay is between \$1102 and \$1290.
8. **Snow.** Based on meteorological data for the past century, a local TV weather forecaster estimates that the region's average winter snowfall is 23", with a margin of error of  $\pm 2$  inches. Assuming he used a 95% confidence interval, how should viewers interpret this news? Comment on each of these statements:
- During 95 of the last 100 winters, the region got between 21" and 25" of snow.
  - There's a 95% chance the region will get between 21" and 25" of snow this winter.
  - There will be between 21" and 25" of snow on the ground for 95% of the winter days.
  - Residents can be 95% sure that the area's average snowfall is between 21" and 25".
  - Residents can be 95% confident that the average snowfall during the last century was between 21" and 25" per winter.
- T** 9. **Pulse rates.** A medical researcher measured the pulse rates (beats per minute) of a sample of randomly selected adults and found the following Student's *t*-based confidence interval:
- Explain carefully what the software output means.
  - What's the margin of error for this interval?
  - If the researcher had calculated a 99% confidence interval, would the margin of error be larger or smaller? Explain.

10. **Crawling.** Data collected by child development scientists produced this confidence interval for the average age (in weeks) at which babies begin to crawl:

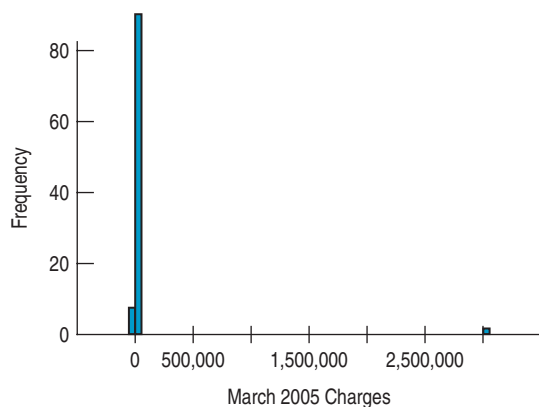
t-Interval for  $\mu$   
(95.00% Confidence):  $29.202 < \mu(\text{age}) < 31.844$

- Explain carefully what the software output means.
  - What is the margin of error for this interval?
  - If the researcher had calculated a 90% confidence interval, would the margin of error be larger or smaller? Explain.
11. **CEO compensation.** A sample of 20 CEOs from the Forbes 500 shows total annual compensations ranging from a minimum of \$0.1 to \$62.24 million. The average for these 20 CEOs is \$7.946 million. Here's a histogram:



Based on these data, a computer program found that a 95% confidence interval for the mean annual compensation of all Forbes 500 CEOs is (1.69, 14.20) \$ million. Why should you be hesitant to trust this confidence interval?

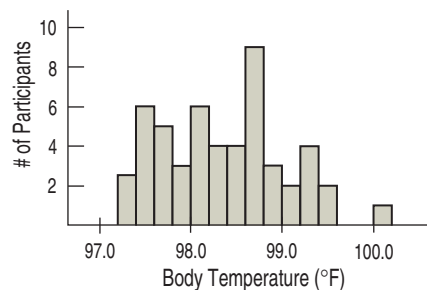
12. **Credit card charges.** A credit card company takes a random sample of 100 cardholders to see how much they charged on their card last month. Here's a histogram.



A computer program found that the resulting 95% confidence interval for the mean amount spent in March 2005 is  $(-\$28366.84, \$90691.49)$ . Explain why the analysts didn't find the confidence interval useful, and explain what went wrong.

- T** 13. **Normal temperature.** The researcher described in Exercise 9 also measured the body temperatures of that randomly selected group of adults. Here are summaries of the data he collected. We wish to estimate the average (or "normal") temperature among the adult population.

Summary	Temperature
Count	52
Mean	98.285
Median	98.200
MidRange	98.600
StdDev	0.6824
Range	2.800
IntQRange	1.050



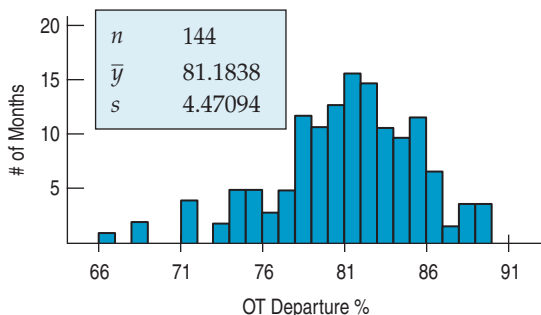
- Check the conditions for creating a  $t$ -interval.
  - Find a 98% confidence interval for mean body temperature.
  - Explain the meaning of that interval.
  - Explain what "98% confidence" means in this context.
  - 98.6°F is commonly assumed to be "normal." Do these data suggest otherwise? Explain.
14. **Parking.** Hoping to lure more shoppers downtown, a city builds a new public parking garage in the central business district. The city plans to pay for the structure through parking fees. During a two-month period (44 weekdays), daily fees collected averaged \$126, with a standard deviation of \$15.
- What assumptions must you make in order to use these statistics for inference?
  - Write a 90% confidence interval for the mean daily income this parking garage will generate.
  - Interpret this confidence interval in context.
  - Explain what "90% confidence" means in this context.
  - The consultant who advised the city on this project predicted that parking revenues would average \$130 per day. Based on your confidence interval, do you think the consultant was correct? Why?
- T** 15. **Normal temperatures, part II.** Consider again the statistics about human body temperature in Exercise 13.
- Would a 90% confidence interval be wider or narrower than the 98% confidence interval you calculated before? Explain. (Don't compute the new interval.)
  - What are the advantages and disadvantages of the 98% confidence interval?
  - If we conduct further research, this time using a sample of 500 adults, how would you expect the 98% confidence interval to change? Explain.
  - How large a sample might allow you to estimate the mean body temperature to within 0.1 degrees with 98% confidence?

16. **Parking II.** Suppose that, for budget planning purposes, the city in Exercise 14 needs a better estimate of the mean daily income from parking fees.
- Someone suggests that the city use its data to create a 95% confidence interval instead of the 90% interval first created. How would this interval be better for the city? (You need not actually create the new interval.)
  - How would the 95% interval be worse for the planners?
  - How could they achieve an interval estimate that would better serve their planning needs?
  - How many days' worth of data should they collect to have 95% confidence of estimating the true mean to within \$3?

17. **Speed of light.** In 1882 Michelson measured the speed of light (usually denoted  $c$  as in Einstein's famous equation  $E = mc^2$ ). His values are in km/sec and have 299,000 subtracted from them. He reported the results of 23 trials with a mean of 756.22 and a standard deviation of 107.12.
- Find a 95% confidence interval for the true speed of light from these statistics.
  - State in words what this interval means. Keep in mind that the speed of light is a physical constant that, as far as we know, has a value that is true throughout the universe.
  - What assumptions must you make in order to use your method?

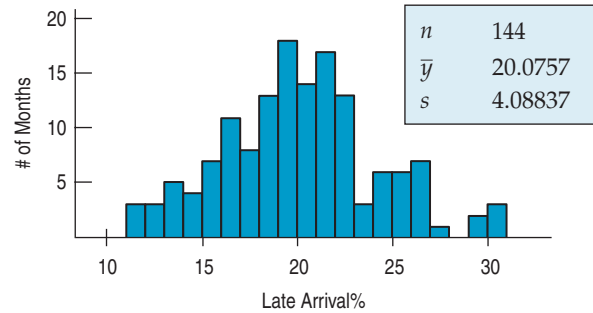
- T** 18. **Better light.** After his first attempt to determine the speed of light (described in Exercise 17), Michelson conducted an "improved" experiment. In 1897 he reported results of 100 trials with a mean of 852.4 and a standard deviation of 79.0.
- What is the standard error of the mean for these data?
  - Without computing it, how would you expect a 95% confidence interval for the second experiment to differ from the confidence interval for the first? Note at least three specific reasons why they might differ, and indicate the ways in which these differences would change the interval.
  - According to Stigler (who reports these values), the true speed of light is 299,710.5 km/sec, corresponding to a value of 710.5 for Michelson's 1897 measurements. What does this indicate about Michelson's two experiments? Explain, using your confidence interval.

- T** 19. **Departures.** What are the chances your flight will leave on time? The U.S. Bureau of Transportation Statistics of the Department of Transportation publishes information about airline performance. Here are a histogram and summary statistics for the percentage of flights departing on time each month from 1995 thru 2006.



There is no evidence of a trend over time. (The correlation of On Time Departure% with time is  $r = -0.016$ .)

- Check the assumptions and conditions for inference.
  - Find a 90% confidence interval for the true percentage of flights that depart on time.
  - Interpret this interval for a traveler planning to fly.
- T** 20. **Late arrivals.** Will your flight get you to your destination on time? The U.S. Bureau of Transportation Statistics reported the percentage of flights that were late each month from 1995 through 2006. Here's a histogram, along with some summary statistics:



We can consider these data to be a representative sample of all months. There is no evidence of a time trend ( $r = -0.07$ ).

- Check the assumptions and conditions for inference about the mean.
  - Find a 99% confidence interval for the true percentage of flights that arrive late.
  - Interpret this interval for a traveler planning to fly.
- T** 21. **For Example, 2nd look.** This chapter's For Examples looked at mirex contamination in farmed salmon. We first found a 95% confidence interval for the mean concentration to be 0.0834 to 0.0992 parts per million. Later we rejected the null hypothesis that the mean did not exceed the EPA's recommended safe level of 0.08 ppm based on a P-value of 0.0027. Explain how these two results are consistent. Your explanation should discuss the confidence level, the P-value, and the decision.
22. **Hot Dogs.** A nutrition lab tested 40 hot dogs to see if their mean sodium content was less than the 325 mg upper limit set by regulations for "reduced sodium" franks. The lab failed to reject the hypothesis that the hot dogs did not meet this requirement, with a P-value of 0.142. A 90% confidence interval estimated the mean sodium content for this kind of hot dog at 317.2 to 326.8 mg. Explain how these two results are consistent. Your explanation should discuss the confidence level, the P-value, and the decision.
23. **Pizza.** A researcher tests whether the mean cholesterol level among those who eat frozen pizza exceeds the value considered to indicate a health risk. She gets a P-value of 0.07. Explain in this context what the "7%" represents.
24. **Golf balls.** The United States Golf Association (USGA) sets performance standards for golf balls. For example, the initial velocity of the ball may not exceed 250 feet per second when measured by an apparatus approved by the USGA. Suppose a manufacturer introduces a new kind of ball and provides a sample for testing. Based on the mean

speed in the test, the USGA comes up with a P-value of 0.34. Explain in this context what the “34%” represents.

25. **TV safety.** The manufacturer of a metal stand for home TV sets must be sure that its product will not fail under the weight of the TV. Since some larger sets weigh nearly 300 pounds, the company’s safety inspectors have set a standard of ensuring that the stands can support an average of over 500 pounds. Their inspectors regularly subject a random sample of the stands to increasing weight until they fail. They test the hypothesis  $H_0: \mu = 500$  against  $H_A: \mu > 500$ , using the level of significance  $\alpha = 0.01$ . If the sample of stands fail to pass this safety test, the inspectors will not certify the product for sale to the general public.
- Is this an upper-tail or lower-tail test? In the context of the problem, why do you think this is important?
  - Explain what will happen if the inspectors commit a Type I error.
  - Explain what will happen if the inspectors commit a Type II error.
26. **Catheters.** During an angiogram, heart problems can be examined via a small tube (a catheter) threaded into the heart from a vein in the patient’s leg. It’s important that the company that manufactures the catheter maintain a diameter of 2.00 mm. (The standard deviation is quite small.) Each day, quality control personnel make several measurements to test  $H_0: \mu = 2.00$  against  $H_A: \mu \neq 2.00$  at a significance level of  $\alpha = 0.05$ . If they discover a problem, they will stop the manufacturing process until it is corrected.
- Is this a one-sided or two-sided test? In the context of the problem, why do you think this is important?
  - Explain in this context what happens if the quality control people commit a Type I error.
  - Explain in this context what happens if the quality control people commit a Type II error.
27. **TV safety revisited.** The manufacturer of the metal TV stands in Exercise 25 is thinking of revising its safety test.
- If the company’s lawyers are worried about being sued for selling an unsafe product, should they increase or decrease the value of  $\alpha$ ? Explain.
  - In this context, what is meant by the power of the test?
  - If the company wants to increase the power of the test, what options does it have? Explain the advantages and disadvantages of each option.
28. **Catheters again.** The catheter company in Exercise 26 is reviewing its testing procedure.
- Suppose the significance level is changed to  $\alpha = 0.01$ . Will the probability of a Type II error increase, decrease, or remain the same?
  - What is meant by the power of the test the company conducts?
  - Suppose the manufacturing process is slipping out of proper adjustment. As the actual mean diameter of the catheters produced gets farther and farther above the desired 2.00 mm, will the power of the quality control test increase, decrease, or remain the same?
  - What could they do to improve the power of the test?
29. **Marriage.** In 1960, census results indicated that the age at which American men first married had a mean of 23.3 years. It is widely suspected that young people today are waiting longer to get married. We want to find out if the mean age of first marriage has increased during the past 40 years.
- Write appropriate hypotheses.
  - We plan to test our hypothesis by selecting a random sample of 40 men who married for the first time last year. Do you think the necessary assumptions for inference are satisfied? Explain.
  - Describe the approximate sampling distribution model for the mean age in such samples.
  - The men in our sample married at an average age of 24.2 years, with a standard deviation of 5.3 years. What’s the P-value for this result?
  - Explain (in context) what this P-value means.
  - What’s your conclusion?
30. **Fuel economy.** A company with a large fleet of cars hopes to keep gasoline costs down and sets a goal of attaining a fleet average of at least 26 miles per gallon. To see if the goal is being met, they check the gasoline usage for 50 company trips chosen at random, finding a mean of 25.02 mpg and a standard deviation of 4.83 mpg. Is this strong evidence that they have failed to attain their fuel economy goal?
- Write appropriate hypotheses.
  - Are the necessary assumptions to make inferences satisfied?
  - Describe the sampling distribution model of mean fuel economy for samples like this.
  - Find the P-value.
  - Explain what the P-value means in this context.
  - State an appropriate conclusion.
- T** 31. **Ruffles.** Students investigating the packaging of potato chips purchased 6 bags of Lay’s Ruffles marked with a net weight of 28.3 grams. They carefully weighed the contents of each bag, recording the following weights (in grams): 29.3, 28.2, 29.1, 28.7, 28.9, 28.5.
- Do these data satisfy the assumptions for inference? Explain.
  - Find the mean and standard deviation of the weights.
  - Create a 95% confidence interval for the mean weight of such bags of chips.
  - Explain in context what your interval means.
  - Comment on the company’s stated net weight of 28.3 grams.
- T** 32. **Doritos.** Some students checked 6 bags of Doritos marked with a net weight of 28.3 grams. They carefully weighed the contents of each bag, recording the following weights (in grams): 29.2, 28.5, 28.7, 28.9, 29.1, 29.5.
- Do these data satisfy the assumptions for inference? Explain.
  - Find the mean and standard deviation of the weights.
  - Create a 95% confidence interval for the mean weight of such bags of chips.
  - Explain in context what your interval means.
  - Comment on the company’s stated net weight of 28.3 grams.
- T** 33. **Popcorn.** Yvon Hopps ran an experiment to test optimum power and time settings for microwave popcorn. His goal was to find a combination of power and time that would deliver high-quality popcorn with less than 10%

of the kernels left unpopped, on average. After experimenting with several bags, he determined that power 9 at 4 minutes was the best combination.

- a) He concluded that this popping method achieved the 10% goal. If it really does not work that well, what kind of error did Hopps make?
- b) To be sure that the method was successful, he popped 8 more bags of popcorn (selected at random) at this setting. All were of high quality, with the following percentages of uncooked popcorn: 7, 13.2, 10, 6, 7.8, 2.8, 2.2, 5.2. Does this provide evidence that he met his goal of an average of no more than 10% uncooked kernels? Explain.

- T 34. Ski wax.** Bjork Larsen was trying to decide whether to use a new racing wax for cross-country skis. He decided that the wax would be worth the price if he could average less than 55 seconds on a course he knew well, so he planned to test the wax by racing on the course 8 times.
- a) Suppose that he eventually decides not to buy the wax, but it really would lower his average time to below 55 seconds. What kind of error would he have made?
  - b) His 8 race times were 56.3, 65.9, 50.5, 52.4, 46.5, 57.8, 52.2, and 43.2 seconds. Should he buy the wax? Explain.

- T 35. Chips Ahoy.** In 1998, as an advertising campaign, the Nabisco Company announced a “1000 Chips Challenge,” claiming that every 18-ounce bag of their Chips Ahoy cookies contained at least 1000 chocolate chips. Dedicated Statistics students at the Air Force Academy (no kidding) purchased some randomly selected bags of cookies, and counted the chocolate chips. Some of their data are given below. (*Chance*, 12, no. 1[1999])

1219 1214 1087 1200 1419 1121 1325 1345  
1244 1258 1356 1132 1191 1270 1295 1135

- a) Check the assumptions and conditions for inference. Comment on any concerns you have.
- b) Create a 95% confidence interval for the average number of chips in bags of Chips Ahoy cookies.
- c) What does this evidence say about Nabisco’s claim? Use your confidence interval to test an appropriate hypothesis and state your conclusion.

- T 36. Yogurt.** *Consumer Reports* tested 14 brands of vanilla yogurt and found these numbers of calories per serving:

160 200 220 230 120 180 140  
130 170 190 80 120 100 170

- a) Check the assumptions and conditions for inference.
- b) Create a 95% confidence interval for the average calorie content of vanilla yogurt.
- c) A diet guide claims that you will get 120 calories from a serving of vanilla yogurt. What does this evidence indicate? Use your confidence interval to test an appropriate hypothesis and state your conclusion.

- T 37. Maze.** Psychology experiments sometimes involve testing the ability of rats to navigate mazes. The mazes are classified according to difficulty, as measured by the mean length of time it takes rats to find the food at the

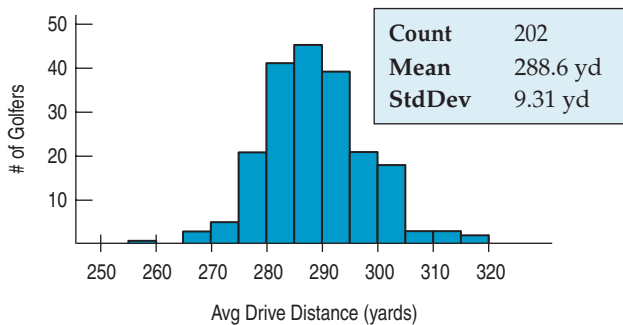
end. One researcher needs a maze that will take rats an average of about one minute to solve. He tests one maze on several rats, collecting the data shown.

Time (sec)	
38.4	57.6
46.2	55.5
62.5	49.5
38.0	40.9
62.8	44.3
33.9	93.8
50.4	47.9
35.0	69.2
52.8	46.2
60.1	56.3
55.1	

- a) Plot the data. Do you think the conditions for inference are satisfied? Explain.
- b) Test the hypothesis that the mean completion time for this maze is 60 seconds. What is your conclusion?
- c) Eliminate the outlier, and test the hypothesis again. What is your conclusion?
- d) Do you think this maze meets the “one-minute average” requirement? Explain.

- 38. Braking.** A tire manufacturer is considering a newly designed tread pattern for its all-weather tires. Tests have indicated that these tires will provide better gas mileage and longer tread life. The last remaining test is for braking effectiveness. The company hopes the tire will allow a car traveling at 60 mph to come to a complete stop within an average of 125 feet after the brakes are applied. They will adopt the new tread pattern unless there is strong evidence that the tires do not meet this objective. The distances (in feet) for 10 stops on a test track were 129, 128, 130, 132, 135, 123, 102, 125, 128, and 130. Should the company adopt the new tread pattern? Test an appropriate hypothesis and state your conclusion. Explain how you dealt with the outlier and why you made the recommendation you did.

- T 39. Driving distance.** How far do professional golfers drive a ball? (For non-golfers, the drive is the shot hit from a tee at the start of a hole and is typically the longest shot.) Here’s a histogram of the average driving distances of the 202 leading professional golfers in 2006 along with summary statistics.

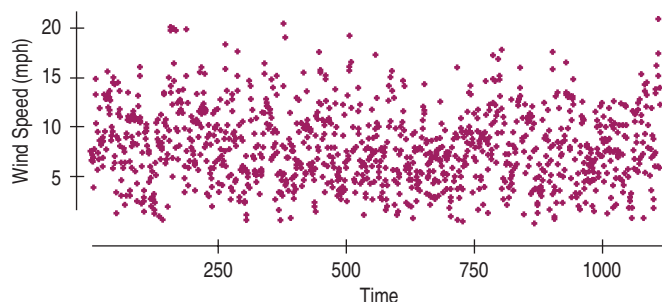
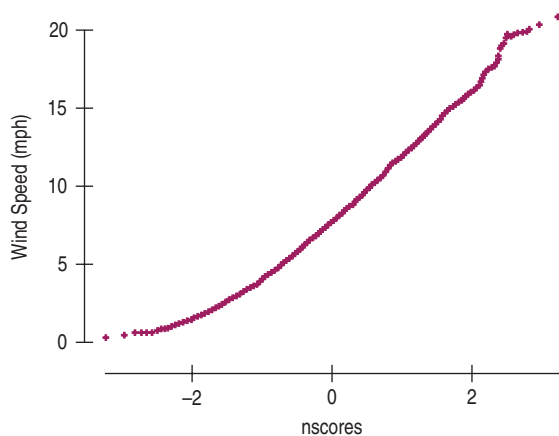
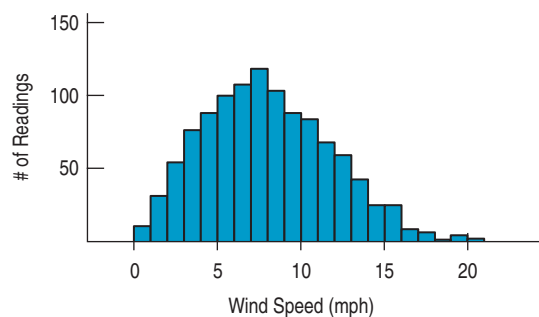


- a) Find a 95% confidence interval for the mean drive distance.
- b) Interpreting this interval raises some problems. Discuss.
- c) The data are the mean driving distance for each golfer. Is that a concern in interpreting the interval? (*Hint*: Review the What Can Go Wrong warnings of Chapter 9. Chapter 9?! Yes, Chapter 9.)

- T 40. Wind power.** Should you generate electricity with your own personal wind turbine? That depends on whether you have enough wind on your site. To produce enough energy, your site should have an annual average wind

speed above 8 miles per hour, according to the Wind Energy Association. One candidate site was monitored for a year, with wind speeds recorded every 6 hours. A total of 1114 readings of wind speed averaged 8.019 mph with a standard deviation of 3.813 mph. You've been asked to make a statistical report to help the landowner decide whether to place a wind turbine at this site.

- a) Discuss the assumptions and conditions for using Student's  $t$  inference methods with these data. Here are some plots that may help you decide whether the methods can be used:



- b) What would you tell the landowner about whether this site is suitable for a small wind turbine? Explain.



## JUST CHECKING Answers

1. Questions on the short form are answered by everyone in the population. This is a census, so means or proportions *are* the true population values. The long forms are given just to a sample of the population. When we estimate parameters from a sample, we use a confidence interval to take sample-to-sample variability into account.
2. They don't know the population standard deviation, so they must use the sample SD as an estimate. The additional uncertainty is taken into account by  $t$ -models.
3. The margin of error for a confidence interval for a mean depends, in part, on the standard error,

$$SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

Since  $n$  is in the denominator, smaller sample sizes lead to larger SEs and correspondingly wider intervals. Long forms returned by one in every six or seven households in a less populous area will be a smaller sample.

4. The critical values for  $t$  with fewer degrees of freedom would be slightly larger. The  $\sqrt{n}$  part of the standard error changes a lot, making the SE much larger. Both would increase the margin of error.
5. The smaller sample is one fourth as large, so the confidence interval would be roughly twice as wide.
6. We expect 95% of such intervals to cover the true value, so 5 of the 100 intervals might be expected to miss.
7. The power would increase if we have a larger sample size.

# Comparing Means



<b>WHO</b>	AA alkaline batteries
<b>WHAT</b>	Length of battery life while playing a CD continuously
<b>UNITS</b>	Minutes
<b>WHY</b>	Class project
<b>WHEN</b>	1998

**A S** **Video: Can Diet Prolong Life?** Watch a video that tells the story of an experiment. We'll analyze the data later in this chapter.

Should you buy generic rather than brand-name batteries? A Statistics student designed a study to test battery life. He wanted to know whether there was any real difference between brand-name batteries and a generic brand. To estimate the difference in mean lifetimes, he kept a battery-powered CD player<sup>1</sup> continuously playing the same CD, with the volume control fixed at 5, and measured the time until no more music was heard through the headphones. (He ran an initial trial to find out approximately how long that would take so that he didn't have to spend the first 3 hours of each run listening to the same CD.) For his trials he used six sets of AA alkaline batteries from two major battery manufacturers: a well-known brand name and a generic brand. He measured the time in minutes until the sound stopped. To account for changes in the CD player's performance over time, he randomized the run order by choosing sets of batteries at random. The table shows his data (times in minutes):

Brand Name	Generic
194.0	190.7
205.5	203.5
199.2	203.5
172.4	206.5
184.0	222.5
169.5	209.4

Studies that compare two groups are common throughout both science and industry. We might want to compare the effects of a new drug with the traditional therapy, the fuel efficiency of two car engine designs, or the sales of new products in two different test cities. In fact, battery manufacturers do research like this on their products and competitors' products themselves.

## Plot the Data

The natural display for comparing two groups is boxplots of the data for the two groups, placed side by side. Although we can't make a confidence interval

<sup>1</sup>Once upon a time, not so very long ago, there were no iPods. At the turn of the century, people actually carried CDs around—and devices to play them. We bet you can find one in your parents' closet.



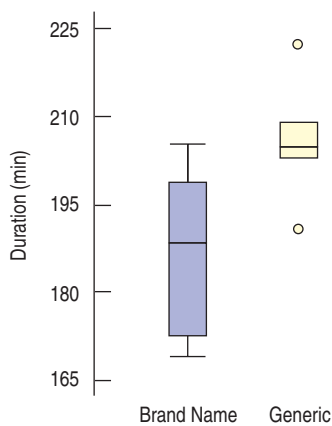


FIGURE 24.1

Boxplots comparing the brand-name and generic batteries suggest a difference in duration.

or test a hypothesis from the boxplots themselves, you should always start with boxplots when comparing groups. Let's look at the boxplots of the battery test data.

It sure looks like the generic batteries lasted longer. And we can see that they were also more consistent. But is the difference large enough to change our battery-buying behavior? Can we be confident that the difference is more than just random fluctuation? That's why we need statistical inference.

The boxplot for the generic data identifies two possible outliers. That's interesting, but with only six measurements in each group, the outlier nomination rule is not very reliable. Both of the extreme values are plausible results, and the range of the generic values is smaller than the range of the brand-name values, even with the outliers. So we're probably better off just leaving these values in the data.

## Comparing Two Means

Comparing two means is not very different from comparing two proportions. In fact, it's not different in concept from any of the methods we've seen. Now, the population model parameter of interest is the difference between the *mean* battery lifetimes of the two brands,  $\mu_1 - \mu_2$ .

The rest is the same as before. The statistic of interest is the difference in the two observed means,  $\bar{y}_1 - \bar{y}_2$ . We'll start with this statistic to build our confidence interval, but we'll need to know its standard deviation and its sampling model. Then we can build confidence intervals and find P-values for hypothesis tests.

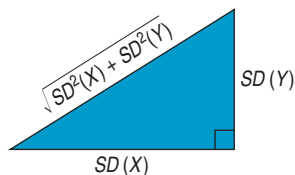
We know that, for independent random variables, the variance of their *difference* is the *sum* of their individual variances,  $Var(Y - X) = Var(Y) + Var(X)$ . To find the standard deviation of the difference between the two independent sample means, we add their variances and then take a square root:

$$\begin{aligned} SD(\bar{y}_1 - \bar{y}_2) &= \sqrt{Var(\bar{y}_1) + Var(\bar{y}_2)} \\ &= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \end{aligned}$$

Of course, we still don't know the true standard deviations of the two groups,  $\sigma_1$  and  $\sigma_2$ , so as usual, we'll use the estimates,  $s_1$  and  $s_2$ . Using the estimates gives us the *standard error*:

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

We'll use the standard error to see how big the difference really is. Because we are working with means and estimating the standard error of their difference using the data, we shouldn't be surprised that the sampling model is a Student's *t*.



The Pythagorean Theorem of Statistics

## FOR EXAMPLE

## Finding the standard error of the difference in independent sample means

Can you tell how much you are eating from how full you are? Or do you need visual cues? Researchers<sup>2</sup> constructed a table with two ordinary 18 oz soup bowls and two identical-looking bowls that had been modified to slowly, imperceptibly, refill as they were emptied. They assigned experiment participants to the bowls randomly and served them tomato soup. Those eating from the ordinary bowls had their bowls refilled by ladle whenever they were one-quarter full. If people judge their portions by internal cues, they should eat about the same amount. How big a difference was there in the amount of soup consumed? The table summarizes their results.

	Ordinary bowl	Refilling bowl
$n$	27	27
$\bar{y}$	8.5 oz	14.7 oz
$s$	6.1 oz	8.4 oz

**Question:** How much variability do we expect in the difference between the two means? Find the standard error.

Participants were randomly assigned to bowls, so the two groups should be independent. It's okay to add variances.

$$SE(\bar{y}_{\text{refill}} - \bar{y}_{\text{ordinary}}) = \sqrt{\frac{s_r^2}{n_r} + \frac{s_o^2}{n_o}} = \sqrt{\frac{8.4^2}{27} + \frac{6.1^2}{27}} = 2.0 \text{ oz.}$$

The confidence interval we build is called a **two-sample  $t$ -interval** (for the difference in means). The corresponding hypothesis test is called a **two-sample  $t$ -test**. The interval looks just like all the others we've seen—the statistic plus or minus an estimated margin of error:

$$(\bar{y}_1 - \bar{y}_2) \pm ME$$

$$\text{where } ME = t^* \times SE(\bar{y}_1 - \bar{y}_2).$$

**z or t?**

If you know  $\sigma$ , use  $z$ .  
(That's rare!)

Whenever you use  $s$   
to estimate  $\sigma$ , use  $t$ .

Compare this formula with the one for the confidence interval for the difference of two proportions we saw in Chapter 22 (page 505). The formulas are almost the same. It's just that here we use a Student's  $t$ -model instead of a Normal model to find the appropriate critical  $t^*$ -value corresponding to our chosen confidence level.

What are we missing? Only the degrees of freedom for the Student's  $t$ -model. Unfortunately, *that* formula is strange.

The deep, dark secret is that the sampling model isn't *really* Student's  $t$ , but only something close. The trick is that by using a special, adjusted degrees-of-freedom value, we can make it so close to a Student's  $t$ -model that nobody can tell the difference. The adjustment formula is straightforward but doesn't help our understanding much, so we leave it to the computer or calculator. (If you are curious and really want to see the formula, look in the footnote.<sup>3</sup>)

<sup>2</sup> Brian Wansink, James E. Painter, and Jill North, "Bottomless Bowls: Why Visual Cues of Portion Size May Influence Intake," *Obesity Research*, Vol. 13, No. 1, January 2005.

<sup>3</sup>

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1} \left(\frac{s_2^2}{n_2}\right)^2}$$

Are you sorry you looked? This formula usually doesn't even give a whole number. If you are using a table, you'll need a whole number, so round down to be safe. If you are using technology, it's even easier. The approximation formulas that computers and calculators use for the Student's  $t$ -distribution deal with degrees of freedom automatically.

### A SAMPLING DISTRIBUTION FOR THE DIFFERENCE BETWEEN TWO MEANS

When the conditions are met, the sampling distribution of the standardized sample difference between the means of two independent groups,

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)},$$

can be modeled by a Student's  $t$ -model with a number of degrees of freedom found with a special formula. We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

## Assumptions and Conditions

Now we've got everything we need. Before we can make a two-sample  $t$ -interval or perform a two-sample  $t$ -test, though, we have to check the assumptions and conditions.

### INDEPENDENCE ASSUMPTION

**Independence Assumption:** The data in each group must be drawn independently and at random from a homogeneous population, or generated by a randomized comparative experiment. We can't expect that the data, taken as one big group, come from a homogeneous population, because that's what we're trying to test. But without randomization of some sort, there are no sampling distribution models and no inference. We can check two conditions:

**Randomization Condition:** Were the data collected with suitable randomization? For surveys, are they a representative random sample? For experiments, was the experiment randomized?

**10% Condition:** We usually don't check this condition for differences of means. We'll check it only if we have a very small population or an extremely large sample. We needn't worry about it at all for randomized experiments.

### NORMAL POPULATION ASSUMPTION

As we did before with Student's  $t$ -models, we should check the assumption that the underlying populations are *each* Normally distributed. We check the . . .

**Nearly Normal Condition:** We must check this for *both* groups; a violation by either one violates the condition. As we saw for single sample means, the Normality Assumption matters most when sample sizes are small. For samples of  $n < 15$  in either group, you should not use these methods if the histogram or Normal probability plot shows severe skewness. For  $n$ 's closer to 40, a mildly skewed histogram is OK, but you should remark on any outliers you find and not work with severely skewed data. When both groups are bigger than 40, the Central Limit Theorem starts to kick in no matter how the data are distributed, so the Nearly Normal Condition for the data matters less. Even in large samples, however, you should still be on the lookout for outliers, extreme skewness, and multiple modes.

### INDEPENDENT GROUPS ASSUMPTION

**Independent Groups Assumption:** To use the two-sample  $t$  methods, the two groups we are comparing must be independent of each other. In fact, this test is

**AS** **Activity: Does Restricting Diet Prolong Life?** This activity lets you construct a confidence interval to compare life spans of rats fed two different diets.

sometimes called the two *independent samples t*-test. No statistical test can verify this assumption. You have to think about how the data were collected. The assumption would be violated, for example, if one group consisted of husbands and the other group their wives. Whatever we measure on couples might naturally be related. Similarly, if we compared subjects' performances before some treatment with their performances afterward, we'd expect a relationship of each "before" measurement with its corresponding "after" measurement. In cases such as these, where the observational units in the two groups are related or matched, *the two-sample methods of this chapter can't be applied*. When this happens, we need a different procedure that we'll see in the next chapter.

## FOR EXAMPLE

### Checking assumptions and conditions

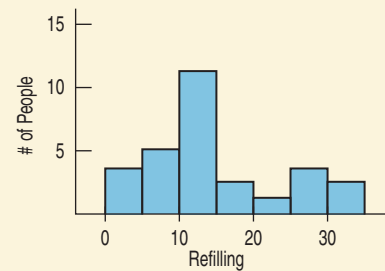
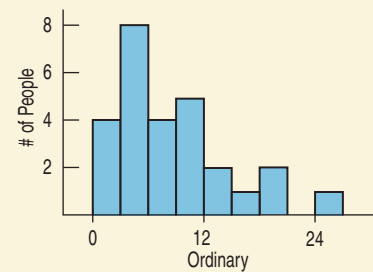
**Recap:** Researchers randomly assigned people to eat soup from one of two bowls: 27 got ordinary bowls that were refilled by ladle, and 27 others bowls that secretly refilled slowly as the people ate.

**Question:** Can the researchers use their data to make inferences about the role of visual cues in determining how much people eat?

- ✓ **Independence Assumption:** The amount consumed by one person should be independent of the amount consumed by others.
- ✓ **Randomization Condition:** Subjects were randomly assigned to the treatments.
- ✓ **Nearly Normal Condition:** The histograms for both groups look unimodal but somewhat skewed to the right. I believe both groups are large enough (27) to allow use of *t*-methods.
- ✓ **Independent Groups Assumption:** Randomization to treatment groups guarantees this.

It's okay to construct a two-sample *t*-interval for the difference in means.

Note: When you check the Nearly Normal Condition it's important that you include the graphs you looked at (histograms or Normal probability plots).



### An Easier Rule?

The formula for the degrees of freedom of the sampling distribution of the difference between two means is long, but the number of degrees of freedom is always at *least* the smaller of the two *n*'s, minus 1. Wouldn't it be easier to just use that value? You could, but *that* approximation can be a poor choice because it can give fewer than *half* the degrees of freedom you're entitled to from the correct formula.

### TWO-SAMPLE *t*-INTERVAL FOR THE DIFFERENCE BETWEEN MEANS

When the conditions are met, we are ready to find the confidence interval for the difference between means of two independent groups,  $\mu_1 - \mu_2$ . The confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{df}^* \times SE(\bar{y}_1 - \bar{y}_2),$$

where the standard error of the difference of the means

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The critical value  $t_{df}^*$  depends on the particular confidence level, *C*, that you specify and on the number of degrees of freedom, which we get from the sample sizes and a special formula.

**FOR EXAMPLE**

**Finding a confidence interval for the difference in sample means**

**Recap:** Researchers studying the role of internal and visual cues in determining how much people eat conducted an experiment in which some people ate soup from bowls that secretly re-filled. The results are summarized in the table.

We've already checked the assumptions and conditions, and have found the standard error for the difference in means to be  $SE(\bar{y}_{refill} - \bar{y}_{ordinary}) = 2.0$  oz.

	Ordinary bowl	Refilling bowl
$n$	27	27
$\bar{y}$	8.5 oz	14.7 oz
$s$	6.1 oz	8.4 oz

**Question:** What does a 95% confidence interval say about the difference in mean amounts eaten?

The observed difference in means is  $\bar{y}_{refill} - \bar{y}_{ordinary} = (14.7 - 8.5) = 6.2$  oz

$$df = 47.46 \quad t_{47.46}^* = 2.011 \text{ (Table gives } t_{45}^* = 2.014.)$$

$$ME = t^* \times SE(\bar{y}_{refill} - \bar{y}_{ordinary}) = 2.011(2.0) = 4.02 \text{ oz}$$

The 95% confidence interval for  $\mu_{refill} - \mu_{ordinary}$  is  $6.2 \pm 4.02$ , or  $(2.18, 10.22)$  oz.

I am 95% confident that people eating from a subtly refilling bowl will eat an average of between 2.18 and 10.22 more ounces of soup than those eating from an ordinary bowl.

**STEP-BY-STEP EXAMPLE**

**A Two-Sample  $t$ -Interval**

Judging from the boxplot, the generic batteries seem to have lasted about 20 minutes longer than the brand-name batteries. Before we change our buying habits, what should we expect to happen with the next batteries we buy?

**Question:** How much longer might the generic batteries last?



**Plan** State what we want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the difference in the means, not the individual group means.

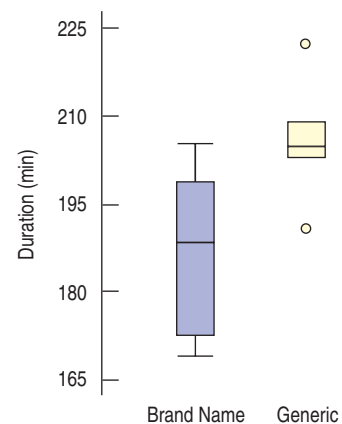
Identify the *population(s)* about which you wish to make statements. We hope to make decisions about purchasing batteries, so we're interested in all the AA batteries of these two brands.

Identify the variables and review the W's.



From the boxplots, it appears our confidence interval should be centered near a difference of 20 minutes. We don't have a lot of intuition about how far the interval should extend on either side of 20.

I have measurements of the lifetimes (in minutes) of 6 sets of generic and 6 sets of brand-name AA batteries from a randomized experiment. I want to find an interval that is likely, with 95% confidence, to contain the true difference  $\mu_G - \mu_B$  between the mean lifetime of the generic AA batteries and the mean lifetime of the brand-name batteries.



**Model** Think about the appropriate assumptions and check the conditions to be sure that a Student's  $t$ -model for the sampling distribution is appropriate.

For very small samples like these, we often don't worry about the 10% Condition.

Make a picture. Boxplots are the display of choice for comparing groups, but now we want to check the *shape* of distribution of each group. Histograms or Normal probability plots do a better job there.

State the sampling distribution model for the statistic. Here the degrees of freedom will come from that messy approximation formula.

Specify your method.

SHOW

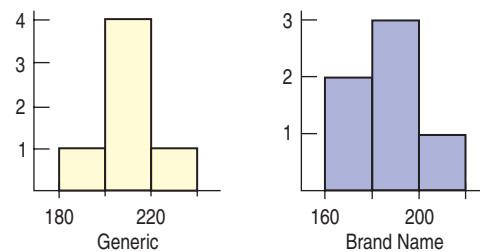
**Mechanics** Construct the confidence interval.

Be sure to include the units along with the statistics. Use meaningful subscripts to identify the groups.

Use the sample standard deviations to find the standard error of the sampling distribution.

We have three choices for degrees of freedom. The best alternative is to let the

- ✓ **Randomization Condition:** The batteries were selected at random from those available for sale. Not exactly an SRS, but a reasonably representative random sample.
- ✓ **Independence Assumption:** The batteries were packaged together, so they may not be independent. For example, a storage problem might affect all the batteries in the same pack. Repeating the study for several different packs of batteries would make the conclusions stronger.
- ✓ **Independent Groups Assumption:** Batteries manufactured by two different companies and purchased in separate packages should be independent.
- ✓ **Nearly Normal Condition:** The samples are small, but the histograms look unimodal and symmetric:



Under these conditions, it's okay to use a Student's  $t$ -model.

I'll use a **two-sample  $t$ -interval**.

$$\begin{aligned} \text{I know } n_G &= 6 & n_B &= 6 \\ \bar{y}_G &= 206.0 \text{ min} & \bar{y}_B &= 187.4 \text{ min} \\ s_G &= 10.3 \text{ min} & s_B &= 14.6 \text{ min} \end{aligned}$$

The groups are independent, so

$$\begin{aligned} SE(\bar{y}_G - \bar{y}_B) &= \sqrt{SE^2(\bar{y}_G) + SE^2(\bar{y}_B)} \\ &= \sqrt{\frac{s_G^2}{n_G} + \frac{s_B^2}{n_B}} \\ &= \sqrt{\frac{10.3^2}{6} + \frac{14.6^2}{6}} \end{aligned}$$

computer or calculator use the approximation formula for df. This gives a fractional degree of freedom (here  $df = 8.98$ ), and technology can find a corresponding critical value. In this case, it is  $t^* = 2.263$ .

Or we could round the approximation formula's df value down to an integer so we can use a  $t$  table. That gives 8 df and a critical value  $t^* = 2.306$ .

The easy rule says to use only  $6 - 1 = 5$  df. That gives a critical value  $t^* = 2.571$ . The corresponding confidence interval is about 14% wider—a high price to pay for a small savings in effort.

$$\begin{aligned} &= \sqrt{\frac{106.09}{6} + \frac{213.16}{6}} \\ &= \sqrt{53.208} \\ &= 7.29 \text{ min.} \end{aligned}$$

$df$  (from technology<sup>4</sup>) = 8.98

The corresponding critical value for a 95% confidence level is  $t^* = 2.263$ .

So the margin of error is

$$\begin{aligned} ME &= t^* \times SE(\bar{y}_G - \bar{y}_B) \\ &= 2.263(7.29) \\ &= 16.50 \text{ min.} \end{aligned}$$

The 95% confidence interval is

$$\begin{aligned} &(206.0 - 187.4) \pm 16.5 \text{ min.} \\ &\text{or } 18.6 \pm 16.5 \text{ min.} \\ &= (2.1, 35.1) \text{ min.} \end{aligned}$$



**Conclusion** Interpret the confidence interval in the proper context.

Less formally, you could say, "I'm 95% confident that generic batteries last an average of 2.1 to 35.1 minutes longer than brand-name batteries."

I am 95% confident that the interval from 2.1 minutes to 35.1 minutes captures the mean amount of time by which generic batteries outlast brand-name batteries for this task. If generic batteries are cheaper, there seems little reason not to use them. If it is more trouble or costs more to buy them, then I'd consider whether the additional performance is worth it.

## Another One Just Like the Other Ones?

**A S**

**Activity: Find Two-Sample  $t$ -Intervals.** Who wants to deal with that ugly df formula? We usually find these intervals with a statistics package. Learn how here.

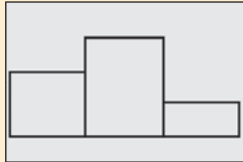
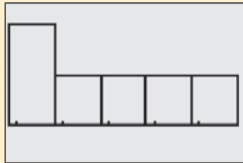
Yes. That's been our point all along. Once again we see a statistic plus or minus the margin of error. And the ME is just a critical value times the standard error. Just look out for that crazy degrees of freedom formula.

### TI Tips

### Creating the confidence interval

If you have been successful using your TI to make confidence intervals for proportions and 1-sample means, then you can probably already use the 2-sample function just fine. But humor us while we do one. Please?

<sup>4</sup> If you try to find the degrees of freedom with that messy approximation formula (We dare you! It's in the footnote on page 562) using the values above, you'll get 8.99. The minor discrepancy is because we rounded the standard deviations to the nearest 10th.



```

2-SampTInt
Inpt: DATA Stats
List1:L1
List2:L2
Freq1:1
Freq2:1
C-Level: .95
↓Pooled: NO Yes

```

```

2-SampTInt
(-35.1, -2.069)
df=8.986279467
x1=187.4333333
x2=206.0166667
sx1=14.6107723
↓sx2=10.3019254

```

### Find a confidence interval for the difference in means, given data from two independent samples.

- Let's do the batteries. Always think about whether the samples are independent. If not, stop right here. These procedures are appropriate only for independent groups.
- Enter the data into two lists.

```

NameBrand in L1:  194.0  205.5  199.2  172.4  184.0  169.5
Generic in L2:   190.7  203.5  203.5  206.5  222.5  209.4

```

- Make histograms of the data to check the Nearly Normal Condition. We see that L1's histogram doesn't look so good. But remember—this is a very small data set. The bars represent only one or two values each. It's not unusual for the histogram to look a little ragged. Try resetting the WINDOW to a range of 160 to 220 with XSc1=20, and Ymax=4. Redraw the GRAPH. Looks better.
- It's your turn to try this. Check L2. Go on, do it.
- Under STAT TESTS choose **0:2-SampTInt**.
- Specify that you are using the Data in L1 and L2, specify 1 for both frequencies, and choose the confidence level you want.
- Pooled? We'll discuss this issue later in the chapter, but the easy advice is: Just Say No.
- To Calculate the interval, you need to scroll down one more line.

Now you have the 95% confidence interval. See **df**? The calculator did that messy degrees of freedom calculation for you. You have to love that!

Notice that the interval bounds are negative. That's because the TI is doing  $\mu_1 - \mu_2$ , and the generic batteries (L2) lasted longer. No harm done—you just need to be careful to interpret that result correctly when you *Tell* what the confidence interval means.

### No data? Find a confidence interval using the sample statistics.

In many situations we don't have the original data, but must work with the summary statistics from the two groups. As we saw in the last chapter, you can still have your TI create the confidence interval with **0:2-SampTInt** by choosing the **Inpt:Stats** option. Enter both means, standard deviations, and sample sizes, then **Calculate**. We show you the details in the next TI Tips.



## JUST CHECKING

Carpal tunnel syndrome (CTS) causes pain and tingling in the hand, sometimes bad enough to keep sufferers awake at night and restrict their daily activities. Researchers studied the effectiveness of two alternative surgical treatments for CTS (Mackenzie, Hainer, and Wheatley, *Annals of Plastic Surgery*, 2000). Patients were randomly assigned to have endoscopic or open-incision surgery. Four weeks later the endoscopic surgery patients demonstrated a mean pinch strength of 9.1 kg compared to 7.6 kg for the open-incision patients.

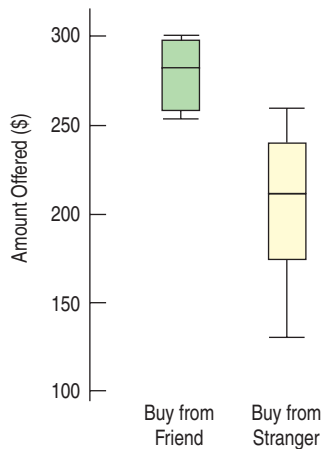
1. Why is the randomization of the patients into the two treatments important?
2. A 95% confidence interval for the difference in mean strength is about (0.04 kg, 2.96 kg). Explain what this interval means.
3. Why might we want to examine such a confidence interval in deciding between these two surgical procedures?
4. Why might you want to see the data before trusting the confidence interval?



## Testing the Difference Between Two Means

If you bought a used camera in good condition from a friend, would you pay the same as you would if you bought the same item from a stranger? A researcher at Cornell University (J. J. Halpern, “The Transaction Index: A Method for Standardizing Comparisons of Transaction Characteristics Across Different Contexts,” *Group Decision and Negotiation*, 6: 557–572) wanted to know how friendship might affect simple sales such as this. She randomly divided subjects into two groups and gave each group descriptions of items they might want to buy. One group was told to imagine buying from a friend whom they expected to see again. The other group was told to imagine buying from a stranger.

Here are the prices they offered for a used camera in good condition:



**WHO** University students  
**WHAT** Prices offered for a used camera  
**UNITS** \$  
**WHY** Study of the effects of friendship on transactions  
**WHEN** 1990s  
**WHERE** U.C. Berkeley

PRICE OFFERED FOR A USED CAMERA (\$)	
Buying from a Friend	Buying from a Stranger
275	260
300	250
260	175
300	130
255	200
275	225
290	240
300	

The researcher who designed this study had a specific concern. Previous theories had doubted that friendship had a measurable effect on pricing. She hoped to find an effect on friendship. This calls for a hypothesis test—in this case a **two-sample *t*-test for the difference between means**.<sup>5</sup>

## A Test for the Difference Between Two Means

**AS** **Activity: The Two-Sample *t*-Test.** How different are beef hot dogs and chicken hot dogs? Test whether measured differences are statistically significant.

You already know enough to construct this test. The test statistic looks just like the others we’ve seen. It finds the difference between the observed group means and compares this with a hypothesized value for that difference. We’ll call that hypothesized difference  $\Delta_0$  (“delta naught”). It’s so common for that hypothesized difference to be zero that we often just assume  $\Delta_0 = 0$ . We then compare the difference in the means with the standard error of that difference. We already know that for a difference between independent means, we can find P-values from a Student’s *t*-model on that same special number of degrees of freedom.

### TWO-SAMPLE *t*-TEST FOR THE DIFFERENCE BETWEEN MEANS

The conditions for the two-sample *t*-test for the difference between the means of two independent groups are the same as for the two-sample *t*-interval. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

<sup>5</sup> Because it is performed so often, this test is usually just called a “two-sample *t*-test.”

**NOTATION ALERT:**

$\Delta_0$ —delta naught—isn't so standard that you can assume everyone will understand it. We use it because it's the Greek letter (good for a parameter) "D" for "difference." You should say "delta naught" rather than "delta zero"—that's standard for parameters associated with null hypotheses.

where the hypothesized difference is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of  $\bar{y}_1 - \bar{y}_2$  is

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

When the conditions are met and the null hypothesis is true, this statistic can be closely modeled by a Student's  $t$ -model with a number of degrees of freedom given by a special formula. We use that model to obtain a P-value.

**STEP-BY-STEP EXAMPLE****A Two-Sample  $t$ -Test for the Difference Between Two Means**

The usual null hypothesis is that there's no difference in means. That's just the right null hypothesis for the camera purchase prices.

**Question:** Is there a difference in the price people would offer a friend rather than a stranger?

**THINK**

**Plan** State what we want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the difference in the means, not the individual group means.

Identify the variables and check the W's.

**Hypotheses** State the null and alternative hypotheses. The research claim is that friendship changes what people are willing to pay.<sup>6</sup> The natural null hypothesis is that friendship makes no difference.

We didn't start with any knowledge of whether friendship might increase or decrease the price, so we choose a two-sided alternative.

**Model** Think about the assumptions and check the conditions. (Note that, because this is a randomized experiment, we haven't sampled at all, so the 10% Condition does not apply.)

I want to know whether people are likely to offer a different amount for a used camera when buying from a friend than when buying from a stranger. I wonder whether the difference between mean amounts is zero. I have bid prices from 8 subjects buying from a friend and 7 buying from a stranger, found in a randomized experiment.

$H_0$ : The difference in mean price offered to friends and the mean price offered to strangers is zero:

$$\mu_F - \mu_S = 0.$$

$H_A$ : The difference in mean prices is not zero:

$$\mu_F - \mu_S \neq 0.$$

- ✓ **Randomization Condition:** The experiment was randomized. Subjects were assigned to treatment groups at random.
- ✓ **Independence Assumption:** This is an experiment, so there is no need for the subjects to be randomly selected from any

<sup>6</sup> This claim is a good example of what is called a "research hypothesis" in many social sciences. The only way to check it is to deny that it's true and see where the resulting null hypothesis leads us.

Make a picture. Boxplots are the display of choice for comparing groups, as seen on page 561. We also want to check the shapes of the distribution. Histograms or Normal probability plots do a better job for that.

State the sampling distribution model.

Specify your method.

**SHOW**

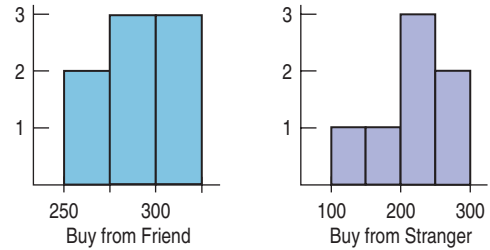
**Mechanics** List the summary statistics. Be sure to use proper notation.

Use the null model to find the P-value. First determine the standard error of the difference between sample means.

Make a picture. Sketch the  $t$ -model centered at the hypothesized difference of zero. Because this is a two-tailed test, shade the region to the right of the observed difference and the corresponding region in the other tail.

particular population. All we need to check is whether they were assigned randomly to treatment groups.

- ✓ **Independent Groups Assumption:** Randomizing the experiment gives independent groups.
- ✓ **Nearly Normal Condition:** Histograms of the two sets of prices are roughly unimodal and symmetric:



The assumptions are reasonable and the conditions are okay, so I'll use a Student's  $t$ -model to perform a **two-sample  $t$ -test**.

From the data:

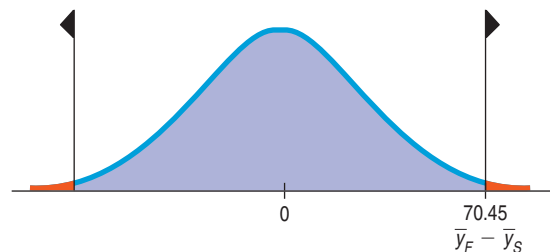
$$\begin{aligned} n_F &= 8 & n_S &= 7 \\ \bar{y}_F &= \$281.88 & \bar{y}_S &= \$211.43 \\ s_F &= \$18.31 & s_S &= \$46.43 \end{aligned}$$

For independent groups,

$$\begin{aligned} SE(\bar{y}_F - \bar{y}_S) &= \sqrt{SE^2(\bar{y}_F) + SE^2(\bar{y}_S)} \\ &= \sqrt{\frac{s_F^2}{n_F} + \frac{s_S^2}{n_S}} \\ &= \sqrt{\frac{18.31^2}{8} + \frac{46.43^2}{7}} \\ &= 18.70 \end{aligned}$$

The observed difference is

$$(\bar{y}_F - \bar{y}_S) = 281.88 - 211.43 = \$70.45$$



Find the  $t$ -value.

A statistics program or graphing calculator finds the P-value using the fractional degrees of freedom from the approximation formula.

$$t = \frac{(\bar{y}_F - \bar{y}_S) - (0)}{SE(\bar{y}_F - \bar{y}_S)} = \frac{70.45}{18.70} = 3.77$$

$df = 7.62$  (from technology)

$$P\text{-value} = 2P(t_{7.62} > 3.77) = 0.006$$



**Conclusion** Link the P-value to your decision about the null hypothesis, and state the conclusion in context.

Be cautious about generalizing to items whose prices are outside the range of those in this study.

If there were no difference in the mean prices, a difference this large would occur only 6 times in 1000. That's too rare to believe, so I reject the null hypothesis and conclude that people are likely to offer a friend more than they'd offer a stranger for a used camera (and possibly for other, similar items).

### TI Tips

### Testing a hypothesis about a difference in means

Now let's use the TI to do a hypothesis test for the difference of two means— independent, of course! (Have we said that enough times yet?)

#### Test a hypothesis when you know the sample statistics.

We'll demonstrate by using the statistics from the camera-pricing example. A sample of 8 people suggested they'd sell the camera to a friend for an average price of \$281.88 with standard deviation \$18.31. An independent sample of 7 other people would charge a stranger an average of \$211.43 with standard deviation \$46.43. Does this represent a significant difference in prices?

- From the **STAT TESTS** menu select **4:2-SampTTest**.
- Specify **Inpt:Stats**, and enter the appropriate sample statistics.
- You have to scroll down to complete the specifications. This is a two-tailed test, so choose alternative **≠μ2**.
- **Pooled?** Just say **No**. (We did promise to explain that and we will, coming up next.)
- Ready ... set ... **Calculate!**

The TI reports a calculated value of  $t = 3.77$  and a P-value of 0.006. It's hard to tell who your real friends are.

#### By now we probably don't have to tell you how to do a 2-SampTTest, starting with data in lists.

So we won't.

```
EDIT CALC TESTS
1:Z-Test...
2:T-Test...
3:2-SampZTest...
4:2-SampTTest...
5:1-PropZTest...
6:2-PropZTest...
7:ZInterval...
```

```
2-SampTTest
Inpt:Data Stats
x1:281.88
Sx1:18.31
n1:8
x2:211.43
Sx2:46.43
n2:7
```

```
2-SampTTest
n1:8
x2:211.43
Sx2:46.43
n2:7
μ1:≠μ2 <μ2 >μ2
Pooled:No Yes
Calculate Draw
```

```
2-SampTTest
μ1≠μ2
t=3.766487374
P=.0059994614
df=7.62304507
x1=281.88
x2=211.43
```



**JUST CHECKING**

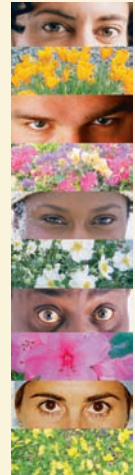
Recall the experiment comparing patients 4 weeks after surgery for carpal tunnel syndrome. The patients who had endoscopic surgery demonstrated a mean pinch strength of 9.1 kg compared to 7.6 kg for the open-incision patients.

5. What hypotheses would you test?
6. The P-value of the test was less than 0.05. State a brief conclusion.
7. The study reports work on 36 “hands,” but there were only 26 patients. In fact, 7 of the endoscopic surgery patients had both hands operated on, as did 3 of the open-incision group. Does this alter your thinking about any of the assumptions? Explain.

**FOR EXAMPLE**

**A two-sample t-test**

Many office “coffee stations” collect voluntary payments for the food consumed. Researchers at the University of Newcastle upon Tyne performed an experiment to see whether the image of eyes watching would change employee behavior.<sup>7</sup> They alternated pictures (seen here) of eyes looking at the viewer with pictures of flowers each week on the cupboard behind the “honesty box.” They measured the consumption of milk to approximate the amount of food consumed and recorded the contributions (in £) each week per liter of milk. The table summarizes their results.



**Question:** Do these results provide evidence that there really is a difference in honesty even when it’s only photographs of eyes that are “watching”?

$$H_0: \mu_{eyes} - \mu_{flowers} = 0$$

$$H_A: \mu_{eyes} - \mu_{flowers} \neq 0$$

	Eyes	Flowers
$n$ (# weeks)	5	5
$\bar{y}$	0.417 £/l	0.151 £/l
$s$	0.1811 £/l	0.067 £/l

- ✓ **Independence Assumption:** The amount paid by one person should be independent of the amount paid by others.
- ✓ **Randomization Condition:** This study was observational. Treatments alternated a week at a time and were applied to the same group of office workers.
- ✓ **Nearly Normal Condition:** I don’t have the data to check, but it seems unlikely there would be outliers in either group. I could be more certain if I could see histograms for both groups.
- ✓ **Independent Groups Assumption:** The same workers were recorded each week, but week-to-week independence is plausible.

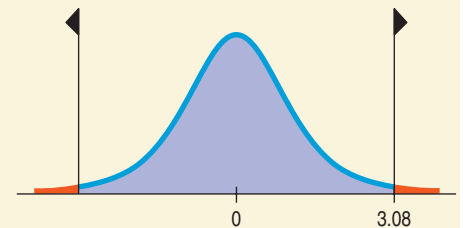
It’s okay to do a two-sample t-test for the difference in means:

$$SE(\bar{y}_{eyes} - \bar{y}_{flowers}) = \sqrt{\frac{s_{eyes}^2}{n_{eyes}} + \frac{s_{flowers}^2}{n_{flowers}}} = \sqrt{\frac{0.1811^2}{5} + \frac{0.067^2}{5}} = 0.0864$$

$$df = 5.07$$

$$t_5 = \frac{(\bar{y}_{eyes} - \bar{y}_{flowers}) - 0}{SE(\bar{y}_{eyes} - \bar{y}_{flowers})} = \frac{0.417 - 0.151}{0.0864} = 3.08$$

$$P(|t_5| > 3.08) = 0.027$$



Assuming the data were free of outliers, the very low P-value leads me to reject the null hypothesis. This study provides evidence that people will leave higher average voluntary payments for food if pictures of eyes are “watching.”

(Note: In Table T we can see that at 5 df,  $t = 3.08$  lies between the critical values for  $P = 0.02$  and  $P = 0.05$ , so we could report  $P < 0.05$ .)

<sup>7</sup>Melissa Bateson, Daniel Nettle, and Gilbert Roberts, “Cues of Being Watched Enhance Cooperation in a Real-World Setting,” *Biol. Lett.* doi:10.1098/rsbl.2006.0509.

## Back into the Pool



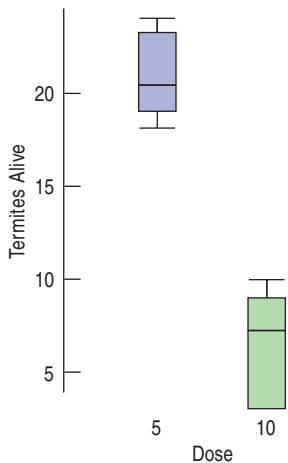
Remember that when we know a proportion, we know its standard deviation. When we tested the null hypothesis that two proportions were equal, that link meant we could assume their variances were equal as well. This led us to pool our data to estimate a standard error for the hypothesis test.

For means, there is also a pooled  $t$ -test. Like the two-proportions  $z$ -test, this test assumes that the variances in the two groups are equal. But be careful: Knowing the mean of some data doesn't tell you anything about their variance. And knowing that two means are equal doesn't say anything about whether their variances are equal. If we were willing to *assume* that their variances are equal, we could pool the data from two groups to estimate the common variance. We'd estimate this pooled variance from the data, so we'd still use a Student's  $t$ -model. This test is called a **pooled  $t$ -test (for the difference between means)**.

Pooled  $t$ -tests have a couple of advantages. They often have a few more degrees of freedom than the corresponding two-sample test and a much simpler degrees of freedom formula. But these advantages come at a price: You have to pool the variances and think about another assumption. The assumption of equal variances is a strong one, is often not true, and is difficult to check. For these reasons, we recommend that you use a two-sample  $t$ -test instead.

The pooled  $t$ -test is the theoretically correct method only when we have a good reason to believe that the variances are equal. And (as we will see shortly) there are times when this makes sense. Keep in mind, however, that it's never wrong *not* to pool.

## \*The Pooled $t$ -Test



Termites cause billions of dollars of damage each year, to homes and other buildings, but some tropical trees seem to be able to resist termite attack. A researcher extracted a compound from the sap of one such tree and tested it by feeding it at two different concentrations to randomly assigned groups of 25 termites.<sup>8</sup> After 5 days, 8 groups fed the lower dose had an average of 20.875 termites alive, with a standard deviation of 2.23. But 6 groups fed the higher dose had an average of only 6.667 termites alive, with a standard deviation of 3.14. Is this a large enough difference to declare the sap compound effective in killing termites? In order to use the pooled  $t$ -test, we must make the **Equal Variance Assumption** that the variances of the two populations from which the samples have been drawn are equal. That is,  $\sigma_1^2 = \sigma_2^2$ . (Of course, we could think about the standard deviations being equal instead.) The corresponding **Similar Spreads Condition** really just consists of looking at the boxplots to check that the spreads are not wildly different. We were going to make boxplots anyway, so there's really nothing new here.

Once we decide to pool, we estimate the common variance by combining numbers we already have:

$$s_{\text{pooled}}^2 = \frac{(8-1)2.23^2 + (6-1)3.14^2}{(8-1) + (6-1)} = 7.01$$

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

(If the two sample sizes are equal, this is just the average of the two variances.)

Now we just substitute this pooled variance in place of each of the variances in the standard error formula.

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{7.01}{8} + \frac{7.01}{6}} = 1.43$$

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} = s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

<sup>8</sup> Adam Messer, Kevin McCormick, Sunjaya, H. H. Hagedorn, Ferny Tumbel, and J. Meinwald, "Defensive role of tropical tree resins: antitermitic sesquiterpenes from Southeast Asian Dipterocarpaceae," *J Chem Ecology*, 16:122, pp. 3333–3352.

The formula for degrees of freedom for the Student's  $t$ -model is simpler, too. It was so complicated for the two-sample  $t$  that we stuck it in a footnote.<sup>9</sup> Now it's just  $df = n_1 + n_2 - 2$ .

$$t = \frac{20.875 - 6.667}{1.43} = 9.935$$

Substitute the pooled- $t$  estimate of the standard error and its degrees of freedom into the steps of the confidence interval or hypothesis test, and you'll be using the pooled- $t$  method. For the termites,  $\bar{y}_1 - \bar{y}_2 = 14.208$ , giving a  $t$ -value = 9.935 with 12 df and a P-value  $\leq 0.0001$ .

Of course, if you decide to use a pooled- $t$  method, you must defend your assumption that the variances of the two groups are equal.

**AS** **Activity: The Pooled  $t$ -Test.** It's those hot dogs again. The same interactive tool can handle a pooled  $t$ -test, too. Take it for a spin here.

### POOLED $t$ -TEST AND CONFIDENCE INTERVAL FOR MEANS

The conditions for the pooled  $t$ -test for the difference between the means of two independent groups (commonly called a "pooled  $t$ -test") are the same as for the two-sample  $t$ -test with the additional assumption that the variances of the two groups are the same. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

where the hypothesized difference,  $\Delta_0$ , is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2)}$$

The standard error of  $\bar{y}_1 - \bar{y}_2$  is

$$SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}} = s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the pooled variance is

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

When the conditions are met and the null hypothesis is true, we can model this statistic's sampling distribution with a Student's  $t$ -model with  $(n_1 - 1) + (n_2 - 1)$  degrees of freedom. We use that model to obtain a P-value for a test or a margin of error for a confidence interval.

The corresponding confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{df}^* \times SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2),$$

where the critical value  $t^*$  depends on the confidence level and is found with  $(n_1 - 1) + (n_2 - 1)$  degrees of freedom.

## Is the Pool All Wet?

We're testing whether the means are equal, so we admit that we don't *know* whether they are equal. Doesn't it seem a bit much to just *assume* that the variances are equal? Well, yes—but there are some special cases to consider. So when *should* you use pooled- $t$  methods rather than two-sample  $t$  methods?

Never.

What, never?

Well, hardly ever.

<sup>9</sup> But not this one. See page 562.

You see, when the variances of the two groups are in fact equal, the two methods give pretty much the same result. (For the termites, the two-sample  $t$  statistic is barely different—9.436 with 8 df—and the P-value is still  $< 0.001$ .) Pooled methods have a small advantage (slightly narrower confidence intervals, slightly more powerful tests) mostly because they usually have a few more degrees of freedom, but the advantage is slight.

When the variances are *not* equal, the pooled methods are just not valid and can give poor results. You have to use the two-sample methods instead.

As the sample sizes get bigger, the advantages that come from a few more degrees of freedom make less and less difference. So the advantage (such as it is) of the pooled method is greatest when the samples are small—just when it's hardest to check the conditions. And the difference in the degrees of freedom is greatest when the variances are not equal—just when you can't use the pooled method anyway. Our advice is to use the two-sample  $t$  methods to compare means.

Pooling may make sense in a randomized comparative experiment. We start by assigning our experimental units to treatments at random, as the experimenter did with the termites. We know that at the start of the experiment each treatment group is a random sample from the same population,<sup>10</sup> so each treatment group begins with the same population variance. In this case, assuming that the variances are equal after we apply the treatment is the same as assuming that the treatment doesn't change the variance. When we test whether the true means are equal, we may be willing to go a bit farther and say that the treatments made no difference *at all*. For example, we might suspect that the treatment is no different from the placebo offered as a control. Then it's not much of a stretch to assume that the variances have remained equal. It's still an assumption, and there are conditions that need to be checked (make the boxplots, make the boxplots, make the boxplots), but at least it's a plausible assumption.

This line of reasoning is important. The methods used to analyze comparative experiments *do* pool variances in exactly this way and defend the pooling with a version of this argument. The chapter on Analysis of Variance on the DVD introduces these methods.

Because the advantages of pooling are small, and you are allowed to pool only rarely (when the Equal Variances Assumption is met), *don't*.

It's never wrong *not* to pool.

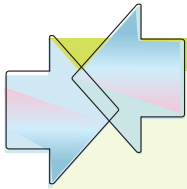
## WHAT CAN GO WRONG?

- ▶ **Watch out for paired data.** The Independent Groups Assumption deserves special attention. If the samples are not independent, you can't use these two-sample methods. This is probably the main thing that can go wrong when using these two-sample methods. The methods of this chapter can be used *only* if the observations in the two groups are *independent*. Matched-pairs designs in which the observations are deliberately related arise often and are important. The next chapter deals with them.
- ▶ **Look at the plots.** The usual (by now) cautions about checking for outliers and non-Normal distributions apply, of course. The simple defense is to make and examine boxplots. You may be surprised how often this simple step saves you from the wrong or even absurd conclusions that can be generated by a single undetected outlier. You don't want to conclude that two methods have very different means just because one observation is atypical.

<sup>10</sup> That is, the population of experimental subjects. Remember that to be valid, experiments do not need a representative sample drawn from a population because we are not trying to estimate a population model parameter.



**Do what we say, not what we do . . .** Precision machines used in industry often have a bewildering number of parameters that have to be set, so experiments are performed in an attempt to try to find the best settings. Such was the case for a hole-punching machine used by a well-known computer manufacturer to make printed circuit boards. The data were analyzed by one of the authors, but because he was in a hurry, he didn't look at the boxplots first and just performed  $t$ -tests on the experimental factors. When he found extremely small  $P$ -values even for factors that made no sense, he plotted the data. Sure enough, there was one observation 1,000,000 times bigger than the others. It turns out that it had been recorded in microns (millionths of an inch), while all the rest were in inches.



## CONNECTIONS

The structure and reasoning of inference methods for comparing two means are very similar to what we used for comparing two proportions. Here we must estimate the standard errors independent of the means, so we use Student's  $t$ -models rather than the Normal.

We first learned about side-by-side boxplots in Chapter 5. There we made general statements about the shape, center, and spread of each group. When we compared groups, we asked whether their centers looked different compared to how spread out the distributions were. Here we've made that kind of thinking precise, with confidence intervals for the difference and tests of whether the means are the same.

We use Student's  $t$  as we did for single sample means, and for the same reasons: We are using standard errors from the data to estimate the standard deviation of the sample statistic. As before, to work with Student's  $t$ -models, we need to check the Nearly Normal Condition. Histograms and Normal probability plots are the best methods for such checks.

As always, we've decided whether a statistic is large by comparing it with its standard error. In this case, our statistic is the difference in means.

We pooled data to find a standard deviation when we tested the hypothesis of equal proportions. For that test, the assumption of equal variances was a consequence of the null hypothesis that the proportions were equal, so it didn't require an extra assumption. When two proportions are equal, so are their variances. But means don't have a linkage with their corresponding variances; so to use pooled- $t$  methods, we must make the additional assumption of equal variances. When we can make this assumption, the pooled variance calculations are very similar to those for proportions, combining the squared deviations of each group from its own mean to find a common variance.



## WHAT HAVE WE LEARNED?

Are the means of two groups the same? If not, how different are they? We've learned to use statistical inference to compare the means of two independent groups.

- ▶ We've seen that confidence intervals and hypothesis tests about the difference between two means, like those for an individual mean, use  $t$ -models.
- ▶ Once again we've seen the importance of checking assumptions that tell us whether our method will work.
- ▶ We've seen that, as when comparing proportions, finding the standard error for the difference in sample means depends on believing that our data come from independent groups. Unlike proportions, however, pooling is usually not the best choice here.
- ▶ And we've seen once again that we can add variances of independent random variables to find the standard deviation of the difference in two independent means.
- ▶ Finally, we've learned that the reasoning of statistical inference remains the same; only the mechanics change.

## Terms

### Two-sample $t$ methods

562. Two-sample  $t$  methods allow us to draw conclusions about the difference between the means of two independent groups. The two-sample methods make relatively few assumptions about the underlying populations, so they are usually the method of choice for comparing two sample means. However, the Student's  $t$ -models are only approximations for their true sampling distribution. To make that approximation work well, the two-sample  $t$  methods have a special rule for estimating degrees of freedom.

### Two-sample $t$ -interval for the difference between means

564. A confidence interval for the difference between the means of two independent groups found as

$$(\bar{y}_1 - \bar{y}_2) \pm t_{df}^* \times SE(\bar{y}_1 - \bar{y}_2)$$

where

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the number of degrees of freedom is given by a special formula (see footnote 3 on page 562).

### Two-sample $t$ -test for the difference between means

569. A hypothesis test for the difference between the means of two independent groups. It tests the null hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference,  $\Delta_0$ , is almost always 0, using the statistic

$$t_{df} = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)},$$

with the number of degrees of freedom given by the special formula.

### Pooling

574. Data from two or more populations may sometimes be combined, or *pooled*, to estimate a statistic (typically a pooled variance) when we are willing to assume that the estimated value is the same in both populations. The resulting larger sample size may lead to an estimate with lower sample variance. However, pooled estimates are appropriate only when the required assumptions are true.

### Pooled- $t$ methods

575. Pooled- $t$  methods provide inferences about the difference between the means of two independent populations under the assumption that both populations have the same standard deviation. When the assumption is justified, pooled- $t$  methods generally produce slightly narrower confidence intervals and more powerful significance tests than two-sample  $t$  methods. When the assumption is not justified, they generally produce worse results—sometimes substantially worse.

We recommend that you use two-sample  $t$  methods instead.

## Skills

THINK

- ▶ Be able to recognize situations in which we want to do inference on the difference between the means of two independent groups.
- ▶ Know how to examine your data for violations of conditions that would make inference about the difference between two population means unwise or invalid.
- ▶ Be able to recognize when a pooled- $t$  procedure might be appropriate and be able to explain why you decided to use a two-sample method anyway.

SHOW

- ▶ Be able to perform a two-sample  $t$ -test using a statistics package or calculator (at least for finding the degrees of freedom).

TELL

- ▶ Be able to interpret a test of the null hypothesis that the means of two independent groups are equal. (If the test is a pooled  $t$ -test, your interpretation should include a defense of your assumption of equal variances.)

## TWO-SAMPLE METHODS ON THE COMPUTER

Here's some typical computer package output with comments:

```

2-Sample t-Test of  $\mu_1 - \mu_2 = 0$  vs  $\neq 0$ 
Difference Between Means = 0.99145299 t-Statistic = 1.540
w/196 df
Fail to reject  $H_0$  at Alpha = 0.05
P = 0.1251

```

Some programs will draw a conclusion about the test. Others just give the P-value and let you decide for yourself.

df found from approximation formula and rounded down. The unrounded value may be given, or may be used to find the P-value.

Many programs give far too many digits. Ignore the excess digits.

Most statistics packages compute the test statistic for you and report a P-value corresponding to that statistic. And, of course, statistics packages make it easy to examine the boxplots and histograms of the two groups, so you have no excuse for skipping this important check.

Some statistics software automatically tries to test whether the variances of the two groups are equal. Some automatically offer both the two-sample-t and pooled-t results. Ignore the test for the variances; it has little power in any situation in which its results could matter. If the pooled and two-sample methods differ in any important way, you should stick with the two-sample method. Most likely, the Equal Variance Assumption needed for the pooled method has failed.

The degrees of freedom approximation usually gives a fractional value. Most packages seem to round the approximate value down to the next smallest integer (although they may actually compute the P-value with the fractional value, gaining a tiny amount of power).

## EXERCISES

- Dogs and calories.** In July 2007, *Consumer Reports* examined the calorie content of two kinds of hot dogs: meat (usually a mixture of pork, turkey, and chicken) and all beef. The researchers purchased samples of several different brands. The meat hot dogs averaged 111.7 calories, compared to 135.4 for the beef hot dogs. A test of the null hypothesis that there's no difference in mean calorie content yields a P-value of 0.124. Would a 95% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  include 0? Explain.
  - The endpoints of this confidence interval are negative numbers. What does that indicate?
  - What does the fact that the confidence interval does not contain 0 indicate?
  - If we use this confidence interval to test the hypothesis that  $\mu_{Meat} - \mu_{Beef} = 0$ , what's the corresponding alpha level?
- Dogs and sodium.** The *Consumer Reports* article described in Exercise 1 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don't differ in the mean amounts of sodium yields a P-value of 0.11. Would a 95% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  include 0? Explain.
- Dogs and fat.** The *Consumer Reports* article described in Exercise 1 also listed the fat content (in grams) for samples of beef and meat hot dogs. The resulting 90% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  is  $(-6.5, -1.4)$ .
  - The endpoints of this confidence interval are negative numbers. What does that indicate?
  - What does the fact that the confidence interval does not contain 0 indicate?
  - If we use this confidence interval to test the hypothesis that  $\mu_{Meat} - \mu_{Beef} = 0$ , what's the corresponding alpha level?
- Washers.** In June 2007, *Consumer Reports* examined top-loading and front-loading washing machines, testing samples of several different brands of each type. One of the variables the article reported was "cycle time", the number of minutes it took each machine to wash a load of clothes. Among the machines rated good to excellent, the 98% confidence interval for the difference in mean cycle time ( $\mu_{Top} - \mu_{Front}$ ) is  $(-40, -22)$ .
  - The endpoints of this confidence interval are negative numbers. What does that indicate?

## TWO-SAMPLE METHODS ON THE COMPUTER

Here's some typical computer package output with comments:

```

2-Sample t-Test of  $\mu_1 - \mu_2 = 0$  vs  $\neq 0$ 

Difference Between Means = 0.99145299 t-Statistic = 1.540
w/196 df
Fail to reject  $H_0$  at Alpha = 0.05
P = 0.1251

```

Some programs will draw a conclusion about the test. Others just give the P-value and let you decide for yourself.

df found from approximation formula and rounded down. The unrounded value may be given, or may be used to find the P-value.

Many programs give far too many digits. Ignore the excess digits.

Most statistics packages compute the test statistic for you and report a P-value corresponding to that statistic. And, of course, statistics packages make it easy to examine the boxplots and histograms of the two groups, so you have no excuse for skipping this important check.

Some statistics software automatically tries to test whether the variances of the two groups are equal. Some automatically offer both the two-sample-t and pooled-t results. Ignore the test for the variances; it has little power in any situation in which its results could matter. If the pooled and two-sample methods differ in any important way, you should stick with the two-sample method. Most likely, the Equal Variance Assumption needed for the pooled method has failed.

The degrees of freedom approximation usually gives a fractional value. Most packages seem to round the approximate value down to the next smallest integer (although they may actually compute the P-value with the fractional value, gaining a tiny amount of power).

## EXERCISES

- Dogs and calories.** In July 2007, *Consumer Reports* examined the calorie content of two kinds of hot dogs: meat (usually a mixture of pork, turkey, and chicken) and all beef. The researchers purchased samples of several different brands. The meat hot dogs averaged 111.7 calories, compared to 135.4 for the beef hot dogs. A test of the null hypothesis that there's no difference in mean calorie content yields a P-value of 0.124. Would a 95% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  include 0? Explain.
  - The endpoints of this confidence interval are negative numbers. What does that indicate?
  - What does the fact that the confidence interval does not contain 0 indicate?
  - If we use this confidence interval to test the hypothesis that  $\mu_{Meat} - \mu_{Beef} = 0$ , what's the corresponding alpha level?
- Dogs and sodium.** The *Consumer Reports* article described in Exercise 1 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don't differ in the mean amounts of sodium yields a P-value of 0.11. Would a 95% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  include 0? Explain.
- Dogs and fat.** The *Consumer Reports* article described in Exercise 1 also listed the fat content (in grams) for samples of beef and meat hot dogs. The resulting 90% confidence interval for  $\mu_{Meat} - \mu_{Beef}$  is  $(-6.5, -1.4)$ .
  - The endpoints of this confidence interval are negative numbers. What does that indicate?
  - What does the fact that the confidence interval does not contain 0 indicate?
  - If we use this confidence interval to test the hypothesis that  $\mu_{Meat} - \mu_{Beef} = 0$ , what's the corresponding alpha level?
- Washers.** In June 2007, *Consumer Reports* examined top-loading and front-loading washing machines, testing samples of several different brands of each type. One of the variables the article reported was "cycle time", the number of minutes it took each machine to wash a load of clothes. Among the machines rated good to excellent, the 98% confidence interval for the difference in mean cycle time ( $\mu_{Top} - \mu_{Front}$ ) is  $(-40, -22)$ .
  - The endpoints of this confidence interval are negative numbers. What does that indicate?

- b) What does the fact that the confidence interval does not contain 0 indicate?
- c) If we use this confidence interval to test the hypothesis that  $\mu_{Top} - \mu_{Front} = 0$ , what's the corresponding alpha level?
5. **Dogs and fat, second helping.** In Exercise 3, we saw a 90% confidence interval of  $(-6.5, -1.4)$  grams for  $\mu_{Meat} - \mu_{Beef}$ , the difference in mean fat content for meat vs. all-beef hot dogs. Explain why you think each of the following statements is true or false:
- If I eat a meat hot dog instead of a beef dog, there's a 90% chance I'll consume less fat.
  - 90% of meat hot dogs have between 1.4 and 6.5 grams less fat than a beef hot dog.
  - I'm 90% confident that meat hot dogs average 1.4–6.5 grams less fat than the beef hot dogs.
  - If I were to get more samples of both kinds of hot dogs, 90% of the time the meat hot dogs would average 1.4–6.5 grams less fat than the beef hot dogs.
  - If I tested many samples, I'd expect about 90% of the resulting confidence intervals to include the true difference in mean fat content between the two kinds of hot dogs.
6. **Second load of wash.** In Exercise 4, we saw a 98% confidence interval of  $(-40, -22)$  minutes for  $\mu_{Top} - \mu_{Front}$ , the difference in time it takes top-loading and front-loading washers to do a load of clothes. Explain why you think each of the following statements is true or false:
- 98% of top loaders are 22 to 40 minutes faster than front loaders.
  - If I choose the laundromat's top loader, there's a 98% chance that my clothes will be done faster than if I had chosen the front loader.
  - If I tried more samples of both kinds of washing machines, in about 98% of these samples I'd expect the top loaders to be an average of 22 to 40 minutes faster.
  - If I tried more samples, I'd expect about 98% of the resulting confidence intervals to include the true difference in mean cycle time for the two types of washing machines.
  - I'm 98% confident that top loaders wash clothes an average of 22 to 40 minutes faster than front-loaders.
7. **Learning math.** The Core Plus Mathematics Project (CPMP) is an innovative approach to teaching Mathematics that engages students in group investigations and mathematical modeling. After field tests in 36 high schools over a three-year period, researchers compared the performances of CPMP students with those taught using a traditional curriculum. In one test, students had to solve applied Algebra problems using calculators. Scores for 320 CPMP students were compared to those of a control group of 273 students in a traditional Math program. Computer software was used to create a confidence interval for the difference in mean scores. (*Journal for Research in Mathematics Education*, 31, no. 3[2000])
- Conf level: 95%    Variable:  $\mu(\text{CPMP}) - \mu(\text{Ctrl})$   
 Inter val: {5.573, 11.427}
- a) What's the margin of error for this confidence interval?

- b) If we had created a 98% CI, would the margin of error be larger or smaller?
- c) Explain what the calculated interval means in context.
- d) Does this result suggest that students who learn Mathematics with CPMP will have significantly higher mean scores in Algebra than those in traditional programs? Explain.

8. **Stereograms.** Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the form of the embedded image affected the time required for subjects to fuse the images. One group of subjects (group NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

2-Sample t-Inter val for  $\mu_1 - \mu_2$   
 Conf level = 90%    df = 70  
 $\mu(\text{NV}) - \mu(\text{VV})$  inter val: {0.55, 5.47}

- Interpret your interval in context.
  - Does it appear that viewing a picture of the image helps people "see" the 3D image in a stereogram?
  - What's the margin of error for this interval?
  - Explain what the 90% confidence level means.
  - Would you expect a 99% confidence level to be wider or narrower? Explain.
  - Might that change your conclusion in part b? Explain.
9. **CPMP, again.** During the study described in Exercise 7, students in both CPMP and traditional classes took another Algebra test that did not allow them to use calculators. The table below shows the results. Are the mean scores of the two groups significantly different?

Math Program	<i>n</i>	Mean	SD
CPMP	312	29.0	18.8
Traditional	265	38.4	16.2

*Performance on Algebraic Symbolic Manipulation Without Use of Calculators*

- Write an appropriate hypothesis.
  - Do you think the assumptions for inference are satisfied? Explain.
  - Here is computer output for this hypothesis test. Explain what the P-value means in this context.
- 2-Sample t-T est of  $\mu_1 - \mu_2 \neq 0$   
 t-Statistic = -6.451 w/574.8761 df  
 P < 0.0001
- d) State a conclusion about the CPMP program.

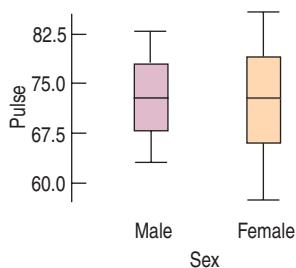
10. **CPMP and word problems.** The study of the new CPMP Mathematics methodology described in Exercise 7 also tested students' abilities to solve word problems. This

table shows how the CPMP and traditional groups performed. What do you conclude?

Math Program	<i>n</i>	Mean	SD
CPMP	320	57.4	32.1
Traditional	273	53.9	28.5

11. **Commuting.** A man who moves to a new city sees that there are two routes he could take to work. A neighbor who has lived there a long time tells him Route A will average 5 minutes faster than Route B. The man decides to experiment. Each day he flips a coin to determine which way to go, driving each route 20 days. He finds that Route A takes an average of 40 minutes, with standard deviation 3 minutes, and Route B takes an average of 43 minutes, with standard deviation 2 minutes. Histograms of travel times for the routes are roughly symmetric and show no outliers.
- Find a 95% confidence interval for the difference in average commuting time for the two routes.
  - Should the man believe the old-timer's claim that he can save an average of 5 minutes a day by always driving Route A? Explain.
12. **Pulse rates.** A researcher wanted to see whether there is a significant difference in resting pulse rates for men and women. The data she collected are displayed in the boxplots and summarized below.

	Sex	
	Male	Female
Count	28	24
Mean	72.75	72.625
Median	73	73
StdDev	5.37225	7.69987
Range	20	29
IQR	9	12.5



- What do the boxplots suggest about differences between male and female pulse rates?
- Is it appropriate to analyze these data using the methods of inference discussed in this chapter? Explain.
- Create a 90% confidence interval for the difference in mean pulse rates.
- Does the confidence interval confirm your answer to part a? Explain.

- T 13. **Cereal.** The data below show the sugar content (as a percentage of weight) of several national brands of children's and adults' cereals. Create and interpret a 95% confidence interval for the difference in mean sugar content. Be sure to check the necessary assumptions and conditions.

**Children's cereals:** 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.6

**Adults' cereals:** 20, 30.2, 2.2, 7.5, 4.4, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4

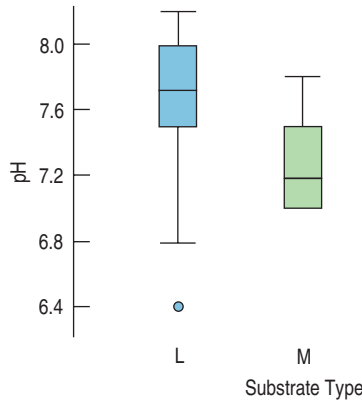
- T 14. **Egyptians.** Some archaeologists theorize that ancient Egyptians interbred with several different immigrant populations over thousands of years. To see if there is any indication of changes in body structure that might have resulted, they measured 30 skulls of male Egyptians dated from 4000 B.C.E and 30 others dated from 200 B.C.E. (A. Thomson and R. Randall-Maciver, *Ancient Races of the Thebaid*, Oxford: Oxford University Press, 1905)
- Are these data appropriate for inference? Explain.
  - Create a 95% confidence interval for the difference in mean skull breadth between these two eras.
  - Do these data provide evidence that the mean breadth of males' skulls changed over this period? Explain.

Maximum Skull Breadth (mm)			
4000 B.C.E.		200 B.C.E.	
131	131	141	131
125	135	141	129
131	132	135	136
119	139	133	131
136	132	131	139
138	126	140	144
139	135	139	141
125	134	140	130
131	128	138	133
134	130	132	138
129	138	134	131
134	128	135	136
126	127	133	132
132	131	136	135
141	124	134	141

- T 15. **Reading.** An educator believes that new reading activities for elementary school children will improve reading comprehension scores. She randomly assigns third graders to an eight-week program in which some will use these activities and others will experience traditional teaching methods. At the end of the experiment, both groups take a reading comprehension exam. Their scores are shown in the back-to-back stem-and-leaf display. Do these results suggest that the new activities are better? Test an appropriate hypothesis and state your conclusion.

New Activities	Control
	1 07
4	2 068
3	3 377
96333	4 1222238
9876432	5 355
721	6 02
1	7
	8 5

- T 16. Streams.** Researchers collected samples of water from streams in the Adirondack Mountains to investigate the effects of acid rain. They measured the pH (acidity) of the water and classified the streams with respect to the kind of substrate (type of rock over which they flow). A lower pH means the water is more acidic. Here is a plot of the pH of the streams by substrate (limestone, mixed, or shale):



Here are selected parts of a software analysis comparing the pH of streams with limestone and shale substrates:

2-Sample t-T est of  $\mu_1 - \mu_2$   
 Difference Between Means = 0.735  
 t-Statistic = 16.30 w/ 133 df  
 $p \leq 0.0001$

- State the null and alternative hypotheses for this test.
  - From the information you have, do the assumptions and conditions appear to be met?
  - What conclusion would you draw?
- T 17. Baseball 2006.** American League baseball teams play their games with the designated hitter rule, meaning that pitchers do not bat. The league believes that replacing the pitcher, traditionally a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. Below are the average numbers of runs scored in American League and National League stadiums for the 2006 season.

American		National	
11.4	9.9	10.5	9.5
10.5	9.7	10.3	9.4
10.4	9.1	10.0	9.1
10.3	9.0	10.0	9.0
10.2	9.0	9.7	9.0
10.0	8.9	9.7	8.9
9.9	8.8	9.6	8.9
		9.5	7.9

- Create an appropriate display of these data. What do you see?
- With a 95% confidence interval, estimate the mean number of runs scored in American League games.
- Coors Field, in Denver, stands a mile above sea level, an altitude far greater than that of any other National League ball park. Some believe that the thinner air makes it harder for pitchers to throw curve balls and easier for

batters to hit the ball a long way. Do you think the 10.5 runs scored per game at Coors is unusual? Explain.  
 d) Explain why you should not use two separate confidence intervals to decide whether the two leagues differ in average number of runs scored.

- 18. Handy.** A factory hiring people to work on an assembly line gives job applicants a test of manual agility. This test counts how many strangely shaped pegs the applicant can fit into matching holes in a one-minute period. The table below summarizes the data by sex of the job applicant. Assume that all conditions necessary for inference are met.

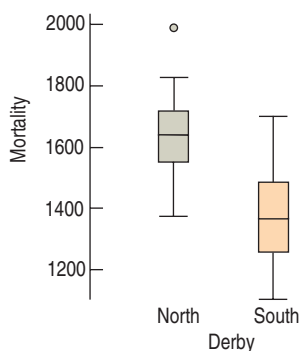
	Male	Female
<b>Number of subjects</b>	50	50
<b>Pegs placed:</b>		
<b>Mean</b>	19.39	17.91
<b>SD</b>	2.52	3.39

- Find 95% confidence intervals for the average number of pegs that males and females can each place.
  - Those intervals overlap. What does this suggest about any sex-based difference in manual agility?
  - Find a 95% confidence interval for the difference in the mean number of pegs that could be placed by men and women.
  - What does this interval suggest about any difference in manual agility between men and women?
  - The two results seem contradictory. Which method is correct: doing two-sample inference or doing one-sample inference twice?
  - Why don't the results agree?
- T 19. Double header 2006.** Do the data in Exercise 17 suggest that the American League's designated hitter rule may lead to more runs?
- Using a 95% confidence interval, estimate the difference between the mean number of runs scored in American and National League games.
  - Interpret your interval.
  - Does that interval suggest that the two leagues may differ in average number of runs scored per game?
- T 20. Hard water.** In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water. The data set also notes, for each town, whether it was south or north of Derby. Is there a significant difference in mortality rates in the two regions? Here are the summary statistics.

Summary of: **mortality**  
 For categories in: **Derby**

Group	Count	Mean	Median	StdDev
<b>North</b>	34	1631.59	1631	138.470
<b>South</b>	27	1388.85	1369	151.114

- Test appropriate hypotheses and state your conclusion.
- On the next page, the boxplots of the two distributions show an outlier among the data north of Derby. What effect might that have had on your test?



- T 21. Job satisfaction.** A company institutes an exercise break for its workers to see if this will improve job satisfaction, as measured by a questionnaire that assesses workers' satisfaction. Scores for 10 randomly selected workers before and after implementation of the exercise program are shown. The company wants to assess the effectiveness of the exercise program. Explain why you can't use the methods discussed in this chapter to do that. (Don't worry, we'll give you another chance to do this the right way.)

Worker Number	Job Satisfaction Index	
	Before	After
1	34	33
2	28	36
3	29	50
4	45	41
5	26	37
6	27	41
7	24	39
8	15	21
9	15	20
10	27	37

- 22. Summer school.** Having done poorly on their math final exams in June, six students repeat the course in summer school, then take another exam in August. If we consider these students representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile?

June	54	49	68	66	62	62
Aug.	50	65	74	64	68	72

- 23. Sex and violence.** In June 2002, the *Journal of Applied Psychology* reported on a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same nine commercials. One of the programs had violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products that were advertised. Here are summaries of the results:

	Program Type		
	Violent	Sexual	Neutral
No. of subjects	108	108	108
Brands recalled			
Mean	2.08	1.71	3.17
SD	1.87	1.76	1.77

- a) Do these results indicate that viewer memory for ads may differ depending on program content? A test of the hypothesis that there is no difference in ad memory between programs with sexual content and those with violent content has a P-value of 0.136. State your conclusion.
- b) Is there evidence that viewer memory for ads may differ between programs with sexual content and those with neutral content? Test an appropriate hypothesis and state your conclusion.
- 24. Ad campaign.** You are a consultant to the marketing department of a business preparing to launch an ad campaign for a new product. The company can afford to run ads during one TV show, and has decided not to sponsor a show with sexual content. You read the study described in Exercise 23, then use a computer to create a confidence interval for the difference in mean number of brand names remembered between the groups watching violent shows and those watching neutral shows.

TWO-SAMPLE T

95% CI FOR  $\mu_{\text{viol}} - \mu_{\text{neut}}$ : [-1.578, 0.602]

- a) At the meeting of the marketing staff, you have to explain what this output means. What will you say?
- b) What advice would you give the company about the upcoming ad campaign?
- 25. Sex and violence II.** In the study described in Exercise 23, the researchers also contacted the subjects again, 24 hours later, and asked them to recall the brands advertised. Results are summarized below.

	Program Type		
	Violent	Sexual	Neutral
No. of subjects	101	106	103
Brands recalled			
Mean	3.02	2.72	4.65
SD	1.61	1.85	1.62

- a) Is there a significant difference in viewers' abilities to remember brands advertised in shows with violent vs. neutral content?
- b) Find a 95% confidence interval for the difference in mean number of brand names remembered between the groups watching shows with sexual content and those watching neutral shows. Interpret your interval in this context.
- 26. Ad recall.** In Exercises 23 and 25, we see the number of advertised brand names people recalled immediately after watching TV shows and 24 hours later. Strangely



enough, it appears that they remembered more about the ads the next day. Should we conclude this is true in general about people’s memory of TV ads?

- a) Suppose one analyst conducts a two-sample hypothesis test to see if memory of brands advertised during violent TV shows is higher 24 hours later. If his P-value is 0.00013, what might he conclude?
  - b) Explain why his procedure was inappropriate. Which of the assumptions for inference was violated?
  - c) How might the design of this experiment have tainted the results?
  - d) Suggest a design that could compare immediate brand-name recall with recall one day later.
27. **Hungry?** Researchers investigated how the size of a bowl affects how much ice cream people tend to scoop when serving themselves.<sup>10</sup> At an “ice cream social,” people were randomly given either a 17 oz or a 34 oz bowl (both large enough that they would not be filled to capacity). They were then invited to scoop as much ice cream as they liked. Did the bowl size change the selected portion size? Here are the summaries:

Small Bowl		Large Bowl	
$n$	26	$n$	22
$\bar{y}$	5.07 oz	$\bar{y}$	6.58 oz
$s$	1.84 oz	$s$	2.91 oz

Test an appropriate hypothesis and state your conclusions. Assume any assumptions and conditions that you cannot test are sufficiently satisfied to proceed.

28. **Thirsty?** Researchers randomly assigned participants either a tall, thin “highball” glass or a short, wide “tumbler,” each of which held 355 ml. Participants were asked to pour a shot (1.5 oz = 44.3 ml) into their glass. Did the shape of the glass make a difference in how much liquid they poured?<sup>11</sup> Here are the summaries:

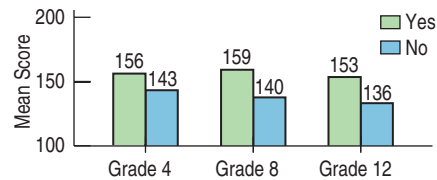
highball		tumbler	
$n$	99	$n$	99
$\bar{y}$	42.2 ml	$\bar{y}$	60.9 ml
$s$	16.2 ml	$s$	17.9 ml

Test an appropriate hypothesis and state your conclusions. Assume any assumptions and conditions that you cannot test are sufficiently satisfied to proceed.

29. **Lower scores?** Newspaper headlines recently announced a decline in science scores among high school seniors. In 2000, a total of 15,109 seniors tested by The National Assessment in Education Program (NAEP)

scored a mean of 147 points. Four years earlier, 7537 seniors had averaged 150 points. The standard error of the difference in the mean scores for the two groups was 1.22.

- a) Have the science scores declined significantly? Cite appropriate statistical evidence to support your conclusion.
  - b) The sample size in 2000 was almost double that in 1996. Does this make the results more convincing or less? Explain.
30. **The Internet.** The NAEP report described in Exercise 29 compared science scores for students who had home Internet access to the scores of those who did not, as shown in the graph. They report that the differences are statistically significant.
- a) Explain what “statistically significant” means in this context.
  - b) If their conclusion is incorrect, which type of error did the researchers commit?
  - c) Does this prove that using the Internet at home can improve a student’s performance in science?



31. **Running heats.** In Olympic running events, preliminary heats are determined by random draw, so we should expect that the abilities of runners in the various heats to be about the same, on average. Here are the times (in seconds) for the 400-m women’s run in the 2004 Olympics in Athens for preliminary heats 2 and 5. Is there any evidence that the mean time to finish is different for randomized heats? Explain. Be sure to include a discussion of assumptions and conditions for your analysis.

Country	Name	Heat	Time
USA	HENNAGAN Monique	2	51.02
BUL	DIMITROVA Mariyana	2	51.29
CHA	NADJINA Kaltouma	2	51.50
JAM	DAVY Nadia	2	52.04
BRA	ALMIRAO Maria Laura	2	52.10
FIN	MYKKANEN Kirsi	2	52.53
CHN	BO Fanfang	2	56.01
BAH	WILLIAMS-DARLING Tonique	5	51.20
BLR	USOVICH Svetlana	5	51.37
UKR	YEFREMOVA Antonina	5	51.53
CMR	NGUIMGO Mireille	5	51.90
JAM	BECKFORD Allison	5	52.85
TOG	THIEBAUD-KANGNI Sandrine	5	52.87
SRI	DHARSHA K V Damayanthi	5	54.58

32. **Swimming heats.** In Exercise 31 we looked at the times in two different heats for the 400-m women’s run from the 2004 Olympics. Unlike track events, swimming heats are *not* determined at random. Instead, swimmers

<sup>10</sup> Brian Wansink, Koert van Ittersum, and James E. Painter, “Ice Cream Illusions: Bowls, Spoons, and Self-Served Portion Sizes,” *Am J Prev Med* 2006.

<sup>11</sup> Brian Wansink and Koert van Ittersum, “Shape of Glass and Amount of Alcohol Poured: Comparative Study of Effect of Practice and Concentration,” *BMJ* 2005;331;1512–1514.

are seeded so that better swimmers are placed in later heats. Here are the times (in seconds) for the women’s 400-m freestyle from heats 2 and 5. Do these results suggest that the mean times of seeded heats are not equal? Explain. Include a discussion of assumptions and conditions for your analysis.

Country	Name	Heat	Time
ARG	BIAGIOLI Cecilia Elizabeth	2	256.42
SLO	CARMAN Anja	2	257.79
CHI	KOBRICH Kristel	2	258.68
MKD	STOJANOVSKA Vesna	2	259.39
JAM	ATKINSON Janelle	2	260.00
NZL	LINTON Rebecca	2	261.58
KOR	HA Eun-Ju	2	261.65
UKR	BERESNYEVA Olga	2	266.30
FRA	MANAUDOU Laure	5	246.76
JPN	YAMADA Sachiko	5	249.10
ROM	PADURARU Simona	5	250.39
GER	STOCKBAUER Hannah	5	250.46
AUS	GRAHAM Elka	5	251.67
CHN	PANG Jiaying	5	251.81
CAN	REIMER Brittany	5	252.33
BRA	FERREIRA Monique	5	253.75

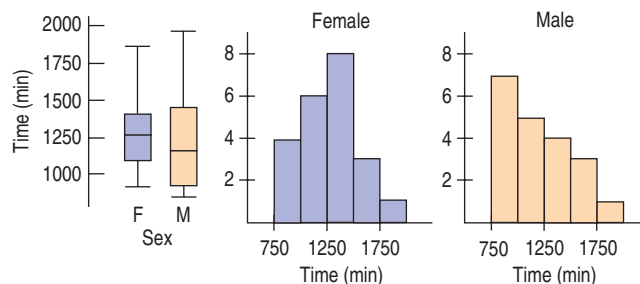
33. **Tees.** Does it matter what kind of tee a golfer places the ball on? The company that manufactures “Stinger” tees claims that the thinner shaft and smaller head will lessen drag, reducing spin and allowing the ball to travel farther. In August 2003, Golf Laboratories, Inc., compared the distance traveled by golf balls hit off regular wooden tees to those hit off Stinger tees. All the balls were struck by the same golf club using a robotic device set to swing the club head at approximately 95 miles per hour. Summary statistics from the test are shown in the table. Assume that 6 balls were hit off each tee and that the data were suitable for inference.

		Total Distance (yards)	Ball Velocity (mph)	Club Velocity (mph)
Regular tee	Avg.	227.17	127.00	96.17
	SD	2.14	0.89	0.41
Stinger tee	Avg.	241.00	128.83	96.17
	SD	2.76	0.41	0.52

Is there evidence that balls hit off the Stinger tees would have a higher initial velocity?

34. **Golf again.** Given the test results on golf tees described in Exercise 33, is there evidence that balls hit off Stinger tees would travel farther? Again, assume that 6 balls were hit off each tee and that the data were suitable for inference.
35. **Crossing Ontario.** Between 1954 and 2003, swimmers have crossed Lake Ontario 43 times. Both women and men have made the crossing. Here are some plots (we’ve

omitted a crossing by Vikki Keith, who swam a round trip—North to South to North—in 3390 minutes):



The summary statistics are:

Summary of Time (min)			
Group	Count	Mean	StdDev
F	22	1271.59	261.111
M	20	1196.75	304.369

How much difference is there between the mean amount of time (in minutes) it would take female and male swimmers to swim the lake?

- a) Construct and interpret a 95% confidence interval for the difference between female and male times.
- b) Comment on the assumptions and conditions.
36. **Music and memory.** Is it a good idea to listen to music when studying for a big test? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, music by Mozart, or no music while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are summary statistics:

	Rap	Mozart	No Music
Count	29	20	13
Mean	10.72	10.00	12.77
SD	3.99	3.19	4.73

- a) Does it appear that it is better to study while listening to Mozart than to rap music? Test an appropriate hypothesis and state your conclusion.
- b) Create a 90% confidence interval for the mean difference in memory score between students who study to Mozart and those who listen to no music at all. Interpret your interval.
37. **Rap.** Using the results of the experiment described in Exercise 36, does it matter whether one listens to rap music while studying, or is it better to study without music at all?
- a) Test an appropriate hypothesis and state your conclusion.
- b) If you concluded there is a difference, estimate the size of that difference with a confidence interval and explain what your interval means.

- T 38. Cuckoos.** Cuckoos lay their eggs in the nests of other (host) birds. The eggs are then adopted and hatched by the host birds. But the potential host birds lay eggs of different sizes. Does the cuckoo change the size of her eggs for different foster species? The numbers in the table are lengths (in mm) of cuckoo eggs found in nests of three different species of other birds. The data are drawn from the work of O.M. Latter in 1902 and were used in a fundamental textbook on statistical quality control by L.H.C. Tippett (1902–1985), one of the pioneers in that field.

CUCKOO EGG LENGTH (MM)		
Foster Parent Species		
Sparrow	Robin	Wagtail
20.85	21.05	21.05
21.65	21.85	21.85
22.05	22.05	21.85
22.85	22.05	21.85
23.05	22.05	22.05
23.05	22.25	22.45
23.05	22.45	22.65
23.05	22.45	23.05
23.45	22.65	23.05
23.85	23.05	23.25
23.85	23.05	23.45
23.85	23.05	24.05
24.05	23.05	24.05
25.05	23.05	24.05
	23.25	24.85
	23.85	

Investigate the question of whether the mean length of cuckoo eggs is the same for different species, and state your conclusion.



### JUST CHECKING Answers

1. Randomization should balance unknown sources of variability in the two groups of patients and helps us believe the two groups are independent.
2. We can be 95% confident that after 4 weeks endoscopic surgery patients will have a mean pinch strength between 0.04 kg and 2.96 kg higher than open-incision patients.
3. The lower bound of this interval is close to 0, so the difference may not be great enough that patients could actually notice the difference. We may want to consider other issues such as cost or risk in making a recommendation about the two surgical procedures.
4. Without data, we can't check the Nearly Normal Condition.
5.  $H_0$ : Mean pinch strength is the same after both surgeries. ( $\mu_E - \mu_O = 0$ )  
 $H_A$ : Mean pinch strength is different after the two surgeries. ( $\mu_E - \mu_O \neq 0$ )
6. With a P-value this low, we reject the null hypothesis. We can conclude that mean pinch strength differs after 4 weeks in patients who undergo endoscopic surgery vs. patients who have open-incision surgery. Results suggest that the endoscopic surgery patients may be stronger, on average.
7. If some patients contributed two hands to the study, then the groups may not be internally independent. It is reasonable to assume that two hands from the same patient might respond in similar ways to similar treatments.

# Paired Samples and Blocks



**WHO** Olympic speed-skaters

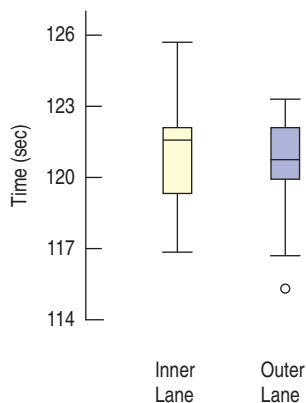
**WHAT** Time for women's 1500 m

**UNITS** Seconds

**WHEN** 2006

**WHERE** Torino, Italy

**WHY** To see whether one lane is faster than the other



**FIGURE 25.1**

Using boxplots to compare times in the inner and outer lanes shows little because it ignores the fact that the skaters raced in pairs.

Speed-skating races are run in pairs. Two skaters start at the same time, one on the inner lane and one on the outer lane. Halfway through the race, they cross over, switching lanes so that each will skate the same distance in each lane. Even though this seems fair, at the 2006 Olympics some fans thought there might have been an advantage to starting on the outside. After all, the winner, Cindy Klassen, started on the outside and skated a remarkable 1.47 seconds faster than the silver medalist.

Here are the data for the women's 1500-m race:

Inner Lane		Outer Lane	
Name	Time	Name	Time
OLTEAN Daniela	129.24	(no competitor)	
ZHANG Xiaolei	125.75	NEMOTO Nami	122.34
ABRAMOVA Yekaterina	121.63	LAMB Maria	122.12
REMPER Shannon	122.24	NOH Seon Yeong	123.35
LEE Ju-Youn	120.85	TIMMER Marianne	120.45
ROKITA Anna Natalia	122.19	MARRA Adelia	123.07
YAKSHINA Valentina	122.15	OPITZ Lucille	122.75
BJELKEVIK Hedvig	122.16	HAUGLI Maren	121.22
ISHINO Eriko	121.85	WOJCICKA Katarzyna	119.96
RANEY Catherine	121.17	BJELKEVIK Annette	121.03
OTSU Hiromi	124.77	LOBYSHEVA Yekaterina	118.87
SIMIONATO Chiara	118.76	JI Jia	121.85
ANSCHUETZ THOMS Daniela	119.74	WANG Fei	120.13
BARYSHEVA Varvara	121.60	van DEUTEKOM Paulien	120.15
GROENEWOLD Renate	119.33	GROVES Kristina	116.74
RODRIGUEZ Jennifer	119.30	NESBITT Christine	119.15
FRIESINGER Anni	117.31	KLASSEN Cindy	115.27
WUST Ireen	116.90	TABATA Maki	120.77

We can view this skating event as an experiment testing whether the lanes were equally fast. Skaters were assigned to lanes randomly. The boxplots of times recorded in the inner and outer lanes (look back a page) don't show much difference. But that's not the right way to compare these times. Conditions can change during the day. The data are recorded for races run two at a time, so the two groups are not independent.

## Paired Data

Data such as these are called **paired**. We have the times for skaters in each lane for each race. The races are run in pairs, so they can't be independent. And since they're not independent, we can't use the two-sample  $t$  methods. Instead, we can focus on the *differences* in times for each racing pair.

Paired data arise in a number of ways. Perhaps the most common way is to compare subjects with themselves before and after a treatment. When pairs arise from an experiment, the pairing is a type of *blocking*. When they arise from an observational study, it is a form of *matching*.

### FOR EXAMPLE

#### Identifying paired data

Do flexible schedules reduce the demand for resources? The Lake County, Illinois, Health Department experimented with a flexible four-day workweek. For a year, the department recorded the mileage driven by 11 field workers on an ordinary five-day workweek. Then it changed to a flexible four-day workweek and recorded mileage for another year.<sup>1</sup> The data are shown.

**Question:** Why are these data paired?

The mileage data are paired because each driver's mileage is measured before and after the change in schedule. I'd expect drivers who drove more than others before the schedule change to continue to drive more afterwards, so the two sets of mileages can't be considered independent.

Name	5-Day mileage	4-Day mileage
Jeff	2798	2914
Betty	7724	6112
Roger	7505	6177
Tom	838	1102
Aimee	4592	3281
Greg	8107	4997
Larry G.	1228	1695
Tad	8718	6606
Larry M.	1097	1063
Leslie	8089	6392
Lee	3807	3362

Pairing isn't a problem; it's an opportunity. If you know the data are paired, you can take advantage of that fact—in fact, you *must* take advantage of it. You *may not* use the two-sample and pooled methods of the previous chapter when the data are paired. Remember: Those methods rely on the Pythagorean Theorem of Statistics, and that requires the two samples be independent. Paired data aren't. There is no test to determine whether the data are paired. You must determine that from understanding how they were collected and what they mean (check the *W*'s).

Once we recognize that the speed-skating data are matched pairs, it makes sense to consider the difference in times for each two-skater race. So we look at the *pairwise* differences:

<sup>1</sup> Charles S. Catlin, "Four-day Work Week Improves Environment," *Journal of Environmental Health*, Denver, 59:7.

**AS** **Activity: Differences in Means of Paired Groups.** Are married couples typically the same age, or do wives tend to be younger than their husbands, on average?

Skating Pair	Inner Time	Outer Time	Inner – Outer
1	129.24		.
2	125.75	122.34	3.41
3	121.63	122.12	-0.49
4	122.24	123.35	-1.11
5	120.85	120.45	0.40
6	122.19	123.07	-0.88
7	122.15	122.75	-0.60
8	122.16	121.22	0.94
9	121.85	119.96	1.89
10	121.17	121.03	0.14
11	124.77	118.87	5.90
12	118.76	121.85	-3.09
13	119.74	120.13	-0.39
14	121.60	120.15	1.45
15	119.33	116.74	2.59
16	119.30	119.15	0.15
17	117.31	115.27	2.04
18	116.90	120.77	-3.87

The first skater raced alone, so we'll omit that race. Because it is the *differences* we care about, we'll treat them as if *they* were the data, ignoring the original two columns. Now that we have only one column of values to consider, we can use a simple one-sample *t*-test. Mechanically, a **paired *t*-test** is just a one-sample *t*-test for the means of these pairwise differences. The sample size is the number of pairs.

So you've already seen the *Show*.

## Assumptions and Conditions



### PAIRED DATA ASSUMPTION

**Paired Data Assumption:** The data must be paired. You can't just decide to pair data when in fact the samples are independent. When you have two groups with the same number of observations, it may be tempting to match them up.

Don't, unless you are prepared to justify your claim that the data are paired.

On the other hand, be sure to recognize paired data when you have them. Remember, two-sample *t* methods aren't valid without independent groups, and paired groups aren't independent. Although this is a strictly required assumption, it is one that can be easy to check if you understand how the data were collected.

### INDEPENDENCE ASSUMPTION

**Independence Assumption:** If the data are paired, the *groups* are not independent. For these methods, it's the *differences* that must be independent of each other. There's no reason to believe that the difference in speeds of one pair of races could affect the difference in speeds for another pair.

**Randomization Condition:** Randomness can arise in many ways. The pairs may be a random sample. In an experiment, the order of the two treatments may be randomly assigned, or the treatments may be randomly assigned to one member of each pair. In a before-and-after study, we may believe that the observed differences are a representative sample from a population of interest. If we have any doubts, we'll need to include a control group to be able to draw conclusions.

**10% of what?**

A fringe benefit of checking the 10% Condition is that it forces us to think about what population we're hoping to make inferences about.

What we want to know usually focuses our attention on where the randomness should be.

In our example, skaters were assigned to the lanes at random.

**10% Condition:** We're thinking of the speed-skating data as an experiment testing the difference between lanes. The 10% Condition doesn't apply to randomized experiments, where no sampling takes place.

**NORMAL POPULATION ASSUMPTION**

We need to assume that the population of *differences* follows a Normal model. We don't need to check the individual groups.

**Nearly Normal Condition:** This condition can be checked with a histogram or Normal probability plot of the *differences*—but not of the individual groups. As with the one-sample *t*-methods, this assumption matters less the more pairs we have to consider. You may be pleasantly surprised when you check this condition. Even if your original measurements are skewed or bimodal, the *differences* may be nearly Normal. After all, the individual who was way out in the tail on an initial measurement is likely to still be out there on the second one, giving a perfectly ordinary difference.

**FOR EXAMPLE****Checking assumptions and conditions**

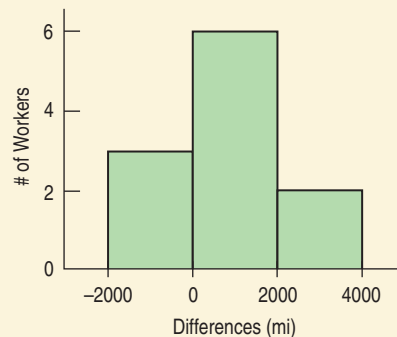
**Recap:** Field workers for a health department compared driving mileage on a five-day work schedule with mileage on a new four-day schedule. To see if the new schedule changed the amount of driving they did, we'll look at paired differences in mileages before and after.

**Question:** Is it okay to use these data to test whether the new schedule changed the amount of driving?

- ✓ **Paired Data Assumption:** The data are paired because each value is the mileage driven by the same person before and after a change in work schedule.
- ✓ **Independence Assumption:** The driving behavior of any individual worker is independent of the others, so the differences are mutually independent.
- ✓ **Randomization Condition:** The mileages are the sums of many individual trips, each of which experienced random events that arose while driving. Repeating the experiment in two new years would give randomly different values.
- ✓ **Nearly Normal Condition:** The histogram of the mileage differences is unimodal and symmetric:

Since the assumptions and conditions are satisfied, it's okay to use paired-*t* methods for these data.

Name	5-Day mileage	4-Day mileage	Difference
Jeff	2798	2914	-116
Betty	7724	6112	1612
Roger	7505	6177	1328
Tom	838	1102	-264
Aimee	4592	3281	1311
Greg	8107	4997	3110
Larry G.	1228	1695	-467
Tad	8718	6606	2112
Larry M.	1097	1063	34
Leslie	8089	6392	1697
Lee	3807	3362	445



The steps in testing a hypothesis for paired differences are very much like the steps for a one-sample *t*-test for a mean.

### THE PAIRED $t$ -TEST

When the conditions are met, we are ready to test whether the mean of paired differences is significantly different from zero. We test the hypothesis

$$H_0: \mu_d = \Delta_0,$$

where the  $d$ 's are the pairwise differences and  $\Delta_0$  is almost always 0.

We use the statistic

$$t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})},$$

where  $\bar{d}$  is the mean of the pairwise differences,  $n$  is the number of *pairs*, and

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

$SE(\bar{d})$  is the ordinary standard error for the mean, applied to the differences.

When the conditions are met and the null hypothesis is true, we can model the sampling distribution of this statistic with a Student's  $t$ -model with  $n - 1$  degrees of freedom, and use that model to obtain a P-value.

### STEP-BY-STEP EXAMPLE

### A Paired $t$ -Test

**Question:** Was there a difference in speeds between the inner and outer speed-skating lanes at the 2006 Winter Olympics?

**THINK**

**Plan** State what we want to know.

Identify the *parameter* we wish to estimate. Here our parameter is the mean difference in race times.

Identify the variables and check the  $W$ 's.

**Hypotheses** State the null and alternative hypotheses.

Although fans suspected one lane was faster, we can't use the data we have to specify the direction of a test. We (and Olympic officials) would be interested in a difference in either direction, so we'd better test a two-sided alternative.

**REALITY CHECK**

The individual differences are all in seconds. We should expect the mean difference to be comparable in magnitude.

**Model** Think about the assumptions and check the conditions.

I want to know whether there really was a difference in the *speeds* of the two lanes for speed skating at the 2006 Olympics. I have data for the women's 1500-m race.

$H_0$ : Neither lane offered an advantage:

$$\mu_d = 0.$$

$H_A$ : The mean difference is different from zero:

$$\mu_d \neq 0.$$

✓ **Independence Assumption:** Each race is independent of the others, so the differences are mutually independent.



State why you think the data are paired. Simply having the same number of individuals in each group and displaying them in side-by-side columns doesn't make them paired.

Think about what we hope to learn and where the randomization comes from. Here, the randomization comes from the racer pairings and lane assignments.

Make a picture—just one. Don't plot separate distributions of the two groups—that entirely misses the pairing. For paired data, it's the Normality of the *differences* that we care about. Treat those paired differences as you would a single variable, and check the Nearly Normal Condition with a histogram or a Normal probability plot.

Specify the sampling distribution model.  
Choose the method.



### Mechanics

$n$  is the number of *pairs*—in this case, the number of races.

$\bar{d}$  is the mean difference.

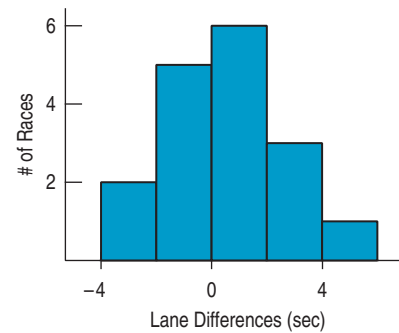
$s_d$  is the standard deviation of the differences.

Find the standard error and the  $t$ -score of the observed mean difference. There is nothing new in the mechanics of the paired- $t$  methods. These are the mechanics of the  $t$ -test for a mean applied to the differences.

Make a picture. Sketch a  $t$ -model centered at the hypothesized mean of 0. Because this is a two-tail test, shade both the region to the right of the observed mean difference of 0.499 seconds and the corresponding region in the lower tail.

Find the P-value, using technology.

- ✓ **Paired Data Assumption:** The data are paired because racers compete in pairs.
- ✓ **Randomization Condition:** Skaters are assigned to lanes at random. Repeating the experiment with different pairings and lane assignments would give randomly different values.
- ✓ **Nearly Normal Condition:** The histogram of the differences is unimodal and symmetric:



The conditions are met, so I'll use a Student's  $t$ -model with  $(n - 1) = 16$  degrees of freedom, and perform a **paired  $t$ -test**.

The data give

$$n = 17 \text{ pairs}$$

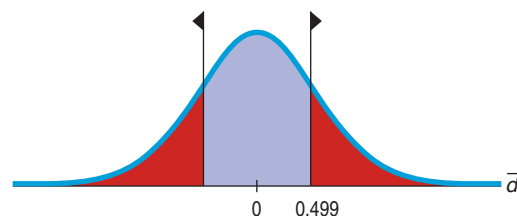
$$\bar{d} = 0.499 \text{ seconds}$$

$$s_d = 2.333 \text{ seconds.}$$

I estimate the standard deviation of  $\bar{d}$  using

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.333}{\sqrt{17}} = 0.5658$$

$$\text{So } t_{16} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{0.499}{0.5658} = 0.882$$



$$\text{P-value} = 2P(t_{16} > 0.882) = 0.39$$

## REALITY CHECK

The mean difference is 0.499 seconds. That may not seem like much, but a smaller difference determined the Silver and Bronze medals. The standard error is about this big, so a  $t$ -value less than 1.0 isn't surprising. Nor is a large  $P$ -value.



**Conclusion** Link the  $P$ -value to your decision about  $H_0$ , and state your conclusion in context.

The  $P$ -value is large. Events that happen more than a third of the time are not remarkable. So, even though there is an observed difference between the lanes, I can't conclude that it isn't due simply to random chance. It appears the fans may have interpreted a random fluctuation in the data as favoring one lane. There's insufficient evidence to declare any lack of fairness.

## FOR EXAMPLE

Doing a paired  $t$ -test

**Recap:** We want to test whether a change from a five-day workweek to a four-day workweek could change the amount driven by field workers of a health department. We've already confirmed that the assumptions and conditions for a paired  $t$ -test are met.

**Question:** Is there evidence that a four-day workweek would change how many miles workers drive?

$H_0$ : The change in the health department workers' schedules didn't change the mean mileage driven; the mean difference is zero:

$$\mu_d = 0.$$

$H_A$ : The mean difference is different from zero:

$$\mu_d \neq 0.$$

The conditions are met, so I'll use a Student's  $t$ -model with  $(n - 1) = 10$  degrees of freedom and perform a **paired  $t$ -test**.

The data give

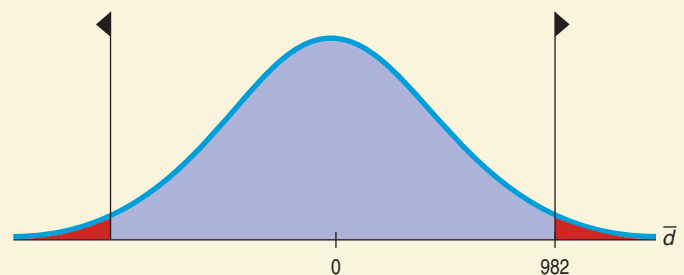
$$n = 11 \text{ pairs}$$

$$\bar{d} = 982 \text{ miles}$$

$$s_d = 1139.6 \text{ miles.}$$

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{1139.6}{\sqrt{11}} = 343.6$$

$$\text{So } t_{10} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{982.0}{343.6} = 2.86$$



$$P\text{-value} = 2P(t_{10} > 2.86) = 0.017$$

The  $P$ -value is small, so I reject the null hypothesis and conclude that the change in workweek did lead to a change in average driving mileage. It appears that changing the work schedule may reduce the mileage driven by workers.

**Note:** We should propose a course of action, but it's hard to tell from the hypothesis test whether the reduction matters. Is the difference in mileage important in the sense of reducing air pollution or costs, or is it merely statistically significant? To help make that decision, we should look at a confidence interval. If the difference in mileage proves to be large in a practical sense, then we might recommend a change in schedule for the rest of the department.

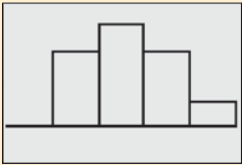
## TI Tips

## Testing a hypothesis with paired data

```
L1-L2→L3
(-116 1612 1328...
```

L1	L2	L3	3
2798	2914	116	
7724	6112	1612	
7505	6177	1328	
838	1102	-264	
4556	2281	1311	
8107	4997	3110	
1228	1695	-467	

L3(1) = -116



```
T-Test
Inpt: Data Stats
μ₀: 0
List: L3
Freq: 1
μ: ≠ μ₀ < μ₀ > μ₀
Calculate Draw
```

```
T-Test
μ ≠ 0
t = 2.85899122
P = .0169862463
x̄ = 982.8181818
Sx = 1140.136116
n = 11
```

Since the inference procedures for matched data are essentially just the one-sample  $t$  procedures, you already know what to do . . . once you have the list of paired differences, that is. That list is not hard to create.

### Test a hypothesis about the mean of paired differences.

- Think: Are the samples independent or paired. Independent? Go back to the last chapter! Paired? Read on.
- Enter the driving data from page 588 into two lists, say *5-Day mileage* in **L1**, *4-Day mileage* in **L2**.
- Create a list of the differences. We want to take each value in **L1**, subtract the corresponding value in **L2**, and store the paired difference in **L3**. The command is **L1-L2 → L3**. (The arrow is the **STO** button.) Now take a look at **L3**. See—it worked!
- Make a histogram of the differences, **L3**, to check the nearly Normal condition. Notice that we do not look at the histograms of the *5-day mileage* or the *4-day mileage*. Those are not the data that we care about now that we are using a paired procedure. Note also that the calculator's first histogram is not close to Normal. More work to do . . .
- As you have seen before, small samples often produce ragged histograms, and these may look very different after a change in bar width. Reset the **WINDOW** to **Xmin=-3000**, **Xmax=4500**, and **Xscl=1500**. The new histogram looks okay.
- Under **STAT TESTS** simply use **2:T-Test**, as you've done before for hypothesis tests about a mean.
- Specify that the hypothesized difference is 0, you're using the **Data** in **L3**, and it's a two-tailed test.
- **Calculate**.

The small P-value shows strong evidence that on average the change in the workweek reduces the number of miles workers drive.

## Confidence Intervals for Matched Pairs

In developed countries, the average age of women is generally higher than that of men. After all, women tend to live longer. But if we look at *married couples*, husbands tend to be slightly older than wives. How much older, on average, are husbands? We have data from a random sample of 200 British couples, the first 7 of which are shown below. Only 170 couples provided ages for both husband and wife, so we can work only with that many pairs. Let's form a confidence interval for the mean difference of husband's and wife's ages for these 170 couples. Here are the first 7 pairs:

**WHO** 170 randomly sampled couples  
**WHAT** Ages  
**UNITS** Years  
**WHEN** Recently  
**WHERE** Britain

Wife's Age	Husband's Age	Difference (husband - wife)
43	49	6
28	25	-3
30	40	10
57	52	-5
52	58	6
27	32	5
52	43	-9
⋮	⋮	⋮

Clearly, these data are paired. The survey selected *couples* at random, not individuals. We're interested in the mean age difference within couples. How would we construct a confidence interval for the true mean difference in ages?

#### PAIRED $t$ -INTERVAL

When the conditions are met, we are ready to find the confidence interval for the mean of the paired differences. The confidence interval is

$$\bar{d} \pm t_{n-1}^* \times SE(\bar{d}),$$

where the standard error of the mean difference is  $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$ .

The critical value  $t^*$  from the Student's  $t$ -model depends on the particular confidence level,  $C$ , that you specify and on the degrees of freedom,  $n - 1$ , which is based on the number of pairs,  $n$ .

Making confidence intervals for matched pairs follows exactly the steps for a one-sample  $t$ -interval.

### STEP-BY-STEP EXAMPLE

#### A Paired $t$ -Interval

**Question:** How big a difference is there, on average, between the ages of husbands and wives?



**Plan** State what we want to know.

Identify the variables and check the W's.

Identify the parameter you wish to estimate. For a paired analysis, the parameter of interest is the mean of the differences. The population of interest is the population of differences.

**Model** Think about the assumptions and check the conditions.

I want to estimate the mean difference in age between husbands and wives. I have a random sample of 200 British couples, 170 of whom provided both ages.

- ✓ **Paired Data Assumption:** The data are paired because they are on members of married couples.
- ✓ **Independence Assumption:** The data are from a randomized survey, so couples should be independent of each other.
- ✓ **Randomization Condition:** These couples were randomly sampled.
- ✓ **10% Condition:** The sample is less than 10% of the population of married couples in Britain.

Make a picture. We focus on the differences, so a histogram or Normal probability plot is best here.

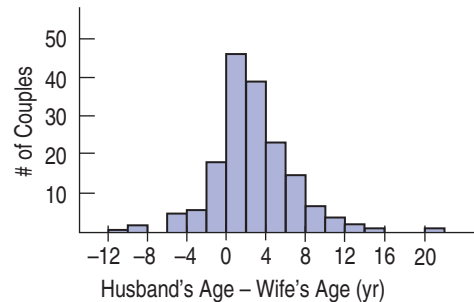
**REALITY CHECK**

The histogram shows husbands are often older than wives (because most of the differences are greater than 0). The mean difference seen here of about 2 years is reasonable.

State the sampling distribution model.

Choose your method.

✓ **Nearly Normal Condition:** The histogram of the husband – wife differences is unimodal and symmetric:



The conditions are met, so I can use a Student's  $t$ -model with  $(n - 1) = 169$  degrees of freedom and find a **paired  $t$ -interval**.

**SHOW**

**Mechanics**

$n$  is the number of *pairs*, here, the number of couples.

$\bar{d}$  is the mean difference.

$s_d$  is the standard deviation of the differences.

Be sure to include the units along with the statistics.

The critical value we need to make a 95% interval comes from a Student's  $t$  table, a computer program, or a calculator.

**REALITY CHECK**

This result makes sense. Our everyday experience confirms that an average age difference of about 2 years is reasonable.

$$n = 170 \text{ couples}$$

$$\bar{d} = 2.2 \text{ years}$$

$$s_d = 4.1 \text{ years}$$

I estimate the standard error of  $\bar{d}$  as

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{4.1}{\sqrt{170}} = 0.31 \text{ years.}$$

The  $df$  for the  $t$ -model is  $n - 1 = 169$ .

The 95% critical value for  $t_{169}$  (from the table) is 1.97.

The margin of error is

$$ME = t_{169}^* \times SE(\bar{d}) = 1.97(0.31) = 0.61$$

So the 95% confidence interval is

$$2.2 \pm 0.6 \text{ years,}$$

or an interval of (1.6, 2.8) years.

**TELL**

**Conclusion** Interpret the confidence interval in context.

I am 95% confident that British husbands are, on average, 1.6 to 2.8 years older than their wives.

## TI Tips

## Creating a confidence interval

```

TIInterval
Inpt:Data Stats
x:2.2
Sx:4.1
n:170
C-Level:.95
Calculate

```

```

TIInterval
(1.5792,2.8208)
x:2.2
Sx:4.1
n:170

```

Now let's get the TI to create a confidence interval for the mean of paired differences.

We'll demonstrate by using the statistics about the ages of the British married couples. (If we had all the data, we could enter that, of course. All 170 couples? Um, no thanks.) The husband in the sample were an average of 2.2 years older than their wives, with a standard deviation of 4.1 years. We've already seen that the data are paired and that a histogram of the differences satisfies the Nearly Normal Condition. (With a sample this large, we could proceed with inference even if we didn't have the actual data and were unable to make the histogram.)

- Once again, we treat the paired differences just like data from one sample. A confidence interval for the mean difference, then, like that for a mean, uses the **STAT TESTS** one-sample procedure **8:TIInterval**.
- Specify **Inpt:Stats**, and enter the statistics for the paired differences.
- **Calculate**.

Done. Finding the interval was the easy part. Now it's time for you to *Tell* what it means. Don't forget to talk about married couples in Britain.

## Effect Size

When we examined the speed-skating times, we failed to reject the null hypothesis, so we couldn't be certain whether there really was a difference between the lanes. Maybe there wasn't any difference, or maybe whatever difference there might have been was just too small to matter at all. Were the fans right to be concerned?

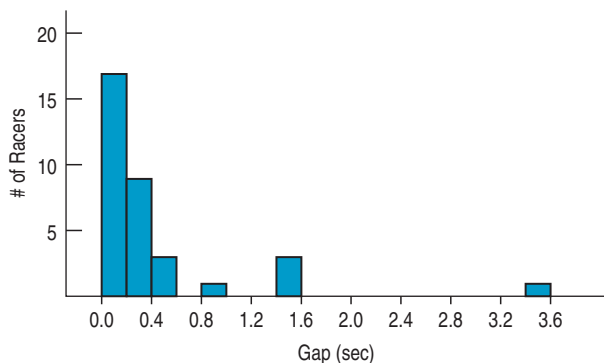
We can't tell from the hypothesis test, but using the same summary statistics, we can find that the corresponding 95% confidence interval for the mean difference is  $(-0.70 < \mu_d < 1.70)$  seconds.

A confidence interval is a good way to get a sense for the size of the effect we're trying to understand. That gives us a plausible range of values for the true mean difference in lane times. If differences of 1.7 seconds were too small to matter

in 1500-m Olympic speed skating, we'd be pretty sure there was no need for concern.

But in fact, except for the Gold – Silver gap, the successive gaps between each skater and the next-faster one were *all* less than the high end of this interval, and most were right around the middle of the interval.

So even though we were unable to discern a real difference, the confidence interval shows that the effects we're considering may be big enough to be important. We may want to continue this investigation by checking out other races on this ice and being alert for possible differences at other venues.



## FOR EXAMPLE

Looking at effect size with a paired- $t$  confidence interval

**Recap:** We know that, on average, the switch from a five-day workweek to a four-day workweek reduced the amount driven by field workers in that Illinois health department. However, finding that there is a significant difference doesn't necessarily mean that difference is meaningful or worthwhile. To assess the size of the effect, we need a confidence interval. We already know the assumptions and conditions are met.

**Question:** By how much, on average, might a change in workweek schedule reduce the amount driven by workers?

$$\begin{aligned}\bar{d} &= 982 \text{ mi} & SE(\bar{d}) &= 343.6 & t_{10}^* &= 2.228 \text{ (for 95\%)} \\ ME &= t_{10}^* \times SE(\bar{d}) & &= 2.228(343.6) & &= 765.54\end{aligned}$$

So the 95% confidence interval for  $\mu_d$  is  $982 \pm 765.54$  or  $(216.46, 1747.54)$  fewer miles.

With 95% confidence, I estimate that by switching to a four-day workweek employees would drive an average of between 216 and 1748 fewer miles per year. With high gas prices, this could save a lot of money.

## Blocking

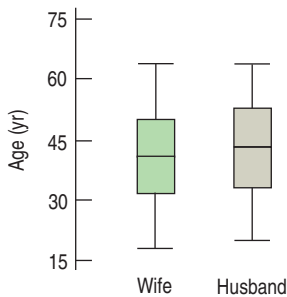


FIGURE 25.2

This display is worthless. It does no good to compare all the wives as a group with all the husbands. We care about the paired differences.

Because the sample of British husbands and wives includes both older and younger couples, there's a lot of variation in the ages of the men and in the ages of the women. In fact, that variation is so great that a boxplot of the two groups would show little difference. But that would be the wrong plot. It's the *difference* we care about. Pairing isolates the extra variation and allows us to focus on the individual differences. In Chapter 13 we saw how we could design an experiment with blocking to isolate the variability between identifiable groups of subjects, allowing us to better see variability among treatment groups due to their response to the treatment. A paired design is an example of blocking.

When we pair, we have roughly half the degrees of freedom of a two-sample test. You may see discussions that suggest that in "choosing" a paired analysis we "give up" these degrees of freedom. This isn't really true, though. If the data are paired, then there never were additional degrees of freedom, and we have no "choice." The fact of the pairing determines how many degrees of freedom are available.

Matching pairs generally removes so much extra variation that it more than compensates for having only half the degrees of freedom. Of course, inappropriate matching when the groups are in fact independent (say, by matching on the first letter of the last name of subjects) would cost degrees of freedom without the benefit of reducing the variance. When you design a study or experiment, you should consider using a paired design if possible.




### JUST CHECKING

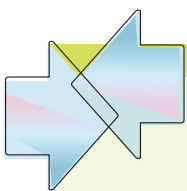
Think about each of the situations described below.

- ▶ Would you use a two-sample  $t$  or paired- $t$  method (or neither)? Why?
  - ▶ Would you perform a hypothesis test or find a confidence interval?
1. Random samples of 50 men and 50 women are asked to imagine buying a birthday present for their best friend. We want to estimate the difference in how much they are willing to spend.
  2. Mothers of twins were surveyed and asked how often in the past month strangers had asked whether the twins were identical.

3. Are parents equally strict with boys and girls? In a random sample of families, researchers asked a brother and sister from each family to rate how strict their parents were.
4. Forty-eight overweight subjects are randomly assigned to either aerobic or stretching exercise programs. They are weighed at the beginning and at the end of the experiment to see how much weight they lost.
  - a) We want to estimate the mean amount of weight lost by those doing aerobic exercise.
  - b) We want to know which program is more effective at reducing weight.
5. Couples at a dance club were separated and each person was asked to rate the band. Do men or women like this band more?

## WHAT CAN GO WRONG?

- ▶ **Don't use a two-sample  $t$ -test when you have paired data.** See the What Can Go Wrong? discussion in Chapter 24.
- ▶ **Don't use a paired- $t$  method when the samples aren't paired.** Just because two groups have the same number of observations doesn't mean they can be paired, even if they are shown side by side in a table. We might have 25 men and 25 women in our study, but they might be completely independent of one another. If they were siblings or spouses, we might consider them paired. Remember that you cannot *choose* which method to use based on your preferences. If the data are from two independent samples, use two-sample  $t$  methods. If the data are from an experiment in which observations were paired, you must use a paired method. If the data are from an observational study, you must be able to defend your decision to use matched pairs or independent groups.
- ▶ **Don't forget outliers.** The outliers we care about now are in the differences. A subject who is extraordinary both before and after a treatment may still have a perfectly typical difference. But one outlying difference can completely distort your conclusions. Be sure to plot the differences (even if you also plot the data).
- ▶ **Don't look for the difference between the means of paired groups with side-by-side boxplots.** The point of the paired analysis is to remove extra variation. The boxplots of each group still contain that variation. Comparing them is likely to be misleading. 



## CONNECTIONS

The most important connection is to the concept of blocking that we first discussed when we considered designed experiments in Chapter 13. Pairing is a basic and very effective form of blocking.

Of course, the details of the mechanics for paired  $t$ -tests and intervals are identical to those for the one-sample  $t$ -methods. Everything we know about those methods applies here.

The connection to the two-sample and pooled methods of the previous chapter is that when the data are naturally paired, those methods are not appropriate because paired data fail the required condition of independence.





## WHAT HAVE WE LEARNED?

When we looked at various ways to design experiments, back in Chapter 13, we saw that pairing can be a very effective strategy. Because pairing can help control variability between individual subjects, paired methods are usually more powerful than methods that compare independent groups. Now we've learned that analyzing data from matched pairs requires different inference procedures.

- ▶ We've learned that paired  $t$ -methods look at pairwise differences. Based on these differences, we test hypotheses and generate confidence intervals. These procedures are mechanically identical to the one-sample  $t$ -methods we saw in Chapter 23.
- ▶ We've also learned to *Think* about the design of the study that collected the data before we proceed with inference. We must be careful to recognize pairing when it is present but not assume it when it is not. Making the correct decision about whether to use independent  $t$ -procedures or paired  $t$ -methods is the first critical step in analyzing the data.

## Terms

### Paired data

588. Data are paired when the observations are collected in pairs or the observations in one group are naturally related to observations in the other. The simplest form of pairing is to measure each subject twice—often before and after a treatment is applied. More sophisticated forms of pairing in experiments are a form of blocking and arise in other contexts. Pairing in observational and survey data is a form of matching.

### Paired $t$ -test

591. A hypothesis test for the mean of the pairwise differences of two groups. It tests the null hypothesis

$$H_0: \mu_d = \Delta_0,$$

where the hypothesized difference is almost always 0, using the statistic

$$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})}$$

with  $n - 1$  degrees of freedom, where  $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$ , and  $n$  is the number of pairs.

### Paired- $t$ confidence interval

595. A confidence interval for the mean of the pairwise differences between paired groups found as

$$\bar{d} \pm t_{n-1}^* \times SE(\bar{d}), \text{ where } SE(\bar{d}) = \frac{s_d}{\sqrt{n}} \text{ and } n \text{ is the number of pairs.}$$

## Skills

THINK

- ▶ Be able to recognize whether a design that compares two groups is paired.

SHOW

- ▶ Be able to find a paired confidence interval, recognizing that it is mechanically equivalent to doing a one-sample  $t$ -interval applied to the differences.
- ▶ Be able to perform a paired  $t$ -test, recognizing that it is mechanically equivalent to a one-sample  $t$ -test applied to the differences.

TELL

- ▶ Be able to interpret a paired  $t$ -test, recognizing that the hypothesis tested is about the mean of the differences between paired values rather than about the differences between the means of two independent groups.
- ▶ Be able to interpret a paired  $t$ -interval, recognizing that it gives an interval for the mean difference in the pairs.

## PAIRED *t* ON THE COMPUTER

Most statistics programs can compute paired-*t* analyses. Some may want you to find the differences yourself and use the one-sample *t* methods. Those that perform the entire procedure will need to know the two variables to compare. The computer, of course, cannot verify that the variables are naturally paired. Most programs will check whether the two variables have the same number of observations, but some stop there, and that can cause trouble. Most programs will automatically omit any pair that is missing a value for either variable (as we did with the British couples). You must look carefully to see whether that has happened.

As we've seen with other inference results, some packages pack a lot of information into a simple table, but you must locate what you want for yourself. Here's a generic example with comments:

Could be called "Matched Pair" or "Paired-*t*" analysis

Individual group means

Mean of the differences and its SE

Paired *t*-statistic

Matched Pairs			
Group 1 Mean	42.9176	t-Ratio	7.151783
Group 2 Mean	40.6824	DF	169
Mean Difference	2.23529	Prob >  t	<0.0001
Std Error	0.31255	Prob > t	<0.0001
Upper 95%	2.85230	Prob < t	1.0000
Lower 95%	1.61829		
N	170		
Correlation	0.93858		

its df

P-values for:  
Two-sided  
One-sided alternatives

Corresponding confidence interval bounds on the mean difference.

Correlation is often reported. Be careful. We have not checked for nonlinearity or outlying pairs. Either could make the correlation meaningless, even though the paired *t* was still appropriate.

Other packages try to be more descriptive. It may be easier to find the results, but you may get less information from the output table.

Groups may have missing values. Only cases with both values present are used in a paired-*t* analysis. You may not learn that from some packages.

Even simple tables can have superfluous numbers such as these.

SD (differences)

SE( $\bar{d}$ )

CI corresponds to specified  $\alpha$ .

Paired T for hAge-wAge				
	N	Mean	Std Dev	SE(Mean)
hAge	199	42.62	11.646	0.8255
wAge	170	40.68	11.414	0.8254
Paired Difference	170	2.235	4.0752	0.31255

$\bar{d}$

95% CI for mean difference: (1.618, 2.852)

T-Test of mean difference = 0 (vs  $\neq$  0): T-Value = 7.1518 P-Value < 0.0001

Some packages let you specify the alternative and report only results for that alternative.

*t*-statistic and its P-value (You may need to calculate  $n_d - 1$  for yourself to get the df.)

Computers make it easy to examine the boxplots of the two groups and the histogram of the differences—both important steps. Some programs offer a scatterplot of the two variables. That can be helpful. In terms of the scatterplot, a paired  $t$ -test is about whether the points tend to be above or below the  $45^\circ$  line  $y = x$ . (Note, though, that pairing says nothing about whether the scatterplot should be straight. That doesn't matter for our  $t$ -methods.)

## EXERCISES

- More eggs?** Can a food additive increase egg production? Agricultural researchers want to design an experiment to find out. They have 100 hens available. They have two kinds of feed: the regular feed and the new feed with the additive. They plan to run their experiment for a month, recording the number of eggs each hen produces.
  - Design an experiment that will require a two-sample  $t$  procedure to analyze the results.
  - Design an experiment that will require a matched-pairs  $t$  procedure to analyze the results.
  - Which experiment would you consider the stronger design? Why?
- MTV.** Some students do homework with the TV on. (Anyone come to mind?) Some researchers want to see if people can work as effectively with as without distraction. The researchers will time some volunteers to see how long it takes them to complete some relatively easy crossword puzzles. During some of the trials, the room will be quiet; during other trials in the same room, a TV will be on, tuned to MTV.
  - Design an experiment that will require a two-sample  $t$  procedure to analyze the results.
  - Design an experiment that will require a matched-pairs  $t$  procedure to analyze the results.
  - Which experiment would you consider the stronger design? Why?
- Sex sells?** Ads for many products use sexual images to try to attract attention to the product. But do these ads bring people's attention to the item that was being advertised? We want to design an experiment to see if the presence of sexual images in an advertisement affects people's ability to remember the product.
  - Describe an experimental design requiring a matched-pairs  $t$  procedure to analyze the results.
  - Describe an experimental design requiring an independent sample procedure to analyze the results.
- Freshman 15?** Many people believe that students gain weight as freshmen. Suppose we plan to conduct a study to see if this is true.
  - Describe a study design that would require a matched-pairs  $t$  procedure to analyze the results.
  - Describe a study design that would require a two-sample  $t$  procedure to analyze the results.
- Women.** Values for the labor force participation rate of women (LFPR) are published by the U.S. Bureau of Labor Statistics. We are interested in whether there was a

difference between female participation in 1968 and 1972, a time of rapid change for women. We check LFPR values for 19 randomly selected cities for 1968 and 1972. Shown below is software output for two possible tests:

Paired t-T est of  $\mu(1 - 2)$   
 Test Ho:  $\mu(1972-1968) = 0$  vs Ha:  $\mu(1972-1968) \neq 0$   
 Mean of Paired Differences = 0.0337  
 t-Statistic = 2.458 w/ 18 df  
 p = 0.0244

2-Sample t-T est of  $\mu_1 - \mu_2$   
 Ho:  $\mu_1 - \mu_2 = 0$  Ha:  $\mu_1 - \mu_2 \neq 0$   
 Test Ho:  $\mu(1972) - \mu(1968) = 0$  vs  
 Ha:  $\mu(1972) - \mu(1968) \neq 0$   
 Difference Between Means = 0.0337  
 t-Statistic = 1.496 w/ 35 df  
 p = 0.1434

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.

- T** 6. **Rain.** Simpson, Alsen, and Eden (*Technometrics* 1975) report the results of trials in which clouds were seeded and the amount of rainfall recorded. The authors report on 26 seeded and 26 unseeded clouds in order of the amount of rainfall, largest amount first. Here are two possible tests to study the question of whether cloud seeding works. Which test is appropriate for these data? Explain your choice. Using the test you select, state your conclusion.

Paired t-T est of  $\mu(1 - 2)$   
 Mean of Paired Differences = -277.39615  
 t-Statistic = -3.641 w/ 25 df  
 p = 0.0012

2-Sample t-T est of  $\mu_1 - \mu_2$   
 Difference Between Means = -277.4  
 t-Statistic = -1.998 w/ 33 df  
 p = 0.0538

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.

- T** 7. **Friday the 13th, I.** In 1993 the *British Medical Journal* published an article titled, "Is Friday the 13th Bad for Your Health?" Researchers in Britain examined how Friday the 13th affects human behavior. One question was

Computers make it easy to examine the boxplots of the two groups and the histogram of the differences—both important steps. Some programs offer a scatterplot of the two variables. That can be helpful. In terms of the scatterplot, a paired  $t$ -test is about whether the points tend to be above or below the  $45^\circ$  line  $y = x$ . (Note, though, that pairing says nothing about whether the scatterplot should be straight. That doesn't matter for our  $t$ -methods.)

## EXERCISES

- More eggs?** Can a food additive increase egg production? Agricultural researchers want to design an experiment to find out. They have 100 hens available. They have two kinds of feed: the regular feed and the new feed with the additive. They plan to run their experiment for a month, recording the number of eggs each hen produces.
  - Design an experiment that will require a two-sample  $t$  procedure to analyze the results.
  - Design an experiment that will require a matched-pairs  $t$  procedure to analyze the results.
  - Which experiment would you consider the stronger design? Why?
- MTV.** Some students do homework with the TV on. (Anyone come to mind?) Some researchers want to see if people can work as effectively with as without distraction. The researchers will time some volunteers to see how long it takes them to complete some relatively easy crossword puzzles. During some of the trials, the room will be quiet; during other trials in the same room, a TV will be on, tuned to MTV.
  - Design an experiment that will require a two-sample  $t$  procedure to analyze the results.
  - Design an experiment that will require a matched-pairs  $t$  procedure to analyze the results.
  - Which experiment would you consider the stronger design? Why?
- Sex sells?** Ads for many products use sexual images to try to attract attention to the product. But do these ads bring people's attention to the item that was being advertised? We want to design an experiment to see if the presence of sexual images in an advertisement affects people's ability to remember the product.
  - Describe an experimental design requiring a matched-pairs  $t$  procedure to analyze the results.
  - Describe an experimental design requiring an independent sample procedure to analyze the results.
- Freshman 15?** Many people believe that students gain weight as freshmen. Suppose we plan to conduct a study to see if this is true.
  - Describe a study design that would require a matched-pairs  $t$  procedure to analyze the results.
  - Describe a study design that would require a two-sample  $t$  procedure to analyze the results.
- Women.** Values for the labor force participation rate of women (LFPR) are published by the U.S. Bureau of Labor Statistics. We are interested in whether there was a

difference between female participation in 1968 and 1972, a time of rapid change for women. We check LFPR values for 19 randomly selected cities for 1968 and 1972. Shown below is software output for two possible tests:

```
Paired t-T est of  $\mu(1 - 2)$ 
Test Ho:  $\mu(1972-1968) = 0$  vs Ha:  $\mu(1972-1968) \neq 0$ 
Mean of Paired Differences = 0.0337
t-Statistic = 2.458 w/18 df
p = 0.0244
```

```
2-Sample t-T est of  $\mu1 - \mu2$ 
Ho:  $\mu1 - \mu2 = 0$  Ha:  $\mu1 - \mu2 \neq 0$ 
Test Ho:  $\mu(1972) - \mu(1968) = 0$  vs
Ha:  $\mu(1972) - \mu(1968) \neq 0$ 
Difference Between Means = 0.0337
t-Statistic = 1.496 w/35 df
p = 0.1434
```

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.

- T** 6. **Rain.** Simpson, Alsen, and Eden (*Technometrics* 1975) report the results of trials in which clouds were seeded and the amount of rainfall recorded. The authors report on 26 seeded and 26 unseeded clouds in order of the amount of rainfall, largest amount first. Here are two possible tests to study the question of whether cloud seeding works. Which test is appropriate for these data? Explain your choice. Using the test you select, state your conclusion.

```
Paired t-T est of  $\mu(1 - 2)$ 
Mean of Paired Differences = -277.39615
t-Statistic = -3.641 w/25 df
p = 0.0012
```

```
2-Sample t-T est of  $\mu1 - \mu2$ 
Difference Between Means = -277.4
t-Statistic = -1.998 w/33 df
p = 0.0538
```

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.

- T** 7. **Friday the 13th, I.** In 1993 the *British Medical Journal* published an article titled, "Is Friday the 13th Bad for Your Health?" Researchers in Britain examined how Friday the 13th affects human behavior. One question was

whether people tend to stay at home more on Friday the 13th. The data below are the number of cars passing Junctions 9 and 10 on the M25 motorway for consecutive Fridays (the 6th and 13th) for five different periods.

Year	Month	6th	13th
1990	July	134,012	132,908
1991	September	133,732	131,843
1991	December	121,139	118,723
1992	March	124,631	120,249
1992	November	117,584	117,263

Here are summaries of two possible analyses:

Paired t-T est of  $\mu[1 - 2] = 0$  vs.  $\mu[1 - 2] > 0$   
 Mean of Paired Differences: 2022.4  
 t-Statistic = 2.9377 w/4 df  
 P = 0.0212

2-Sample t-T est of  $\mu_1 = \mu_2$  vs.  $\mu_1 > \mu_2$   
 Difference Between Means: 2022.4  
 t-Statistic = 0.4273 w/7.998 df  
 P = 0.3402

- Which of the tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.
- Are the assumptions and conditions for inference met?

- T** 8. **Friday the 13th, II:** The researchers in Exercise 7 also examined the number of people admitted to emergency rooms for vehicular accidents on 12 Friday evenings (6 each on the 6th and 13th).

Year	Month	6th	13th
1989	October	9	13
1990	July	6	12
1991	September	11	14
1991	December	11	10
1992	March	3	4
1992	November	5	12

Based on these data, is there evidence that more people are admitted, on average, on Friday the 13th? Here are two possible analyses of the data:

Paired t-T est of  $\mu[1 - 2] = 0$  vs.  $\mu[1 - 2] < 0$   
 Mean of Paired Differences = 3.333  
 t-Statistic = 2.7116 w/5 df  
 P = 0.0211

2-Sample t-T est of  $\mu_1 = \mu_2$  vs.  $\mu_1 < \mu_2$   
 Difference Between Means = 3.333  
 t-Statistic = 1.6644 w/9.940 df  
 P = 0.0636

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.
- Are the assumptions and conditions for inference met?

9. **Online insurance I.** After seeing countless commercials claiming one can get cheaper car insurance from an online company, a local insurance agent was concerned that he might lose some customers. To investigate, he randomly selected profiles (type of car, coverage, driving record, etc.) for 10 of his clients and checked online price quotes for their policies. The comparisons are shown in the table below. His statistical software produced the following summaries (where  $PriceDiff = Local - Online$ ):

Variable	Count	Mean	StdDev
Local	10	799.200	229.281
Online	10	753.300	256.267
PriceDiff	10	45.9000	175.663

Local	Online	PriceDiff
568	391	177
872	602	270
451	488	-37
1229	903	326
605	677	-72
1021	1270	-249
783	703	80
844	789	55
907	1008	-101
712	702	10

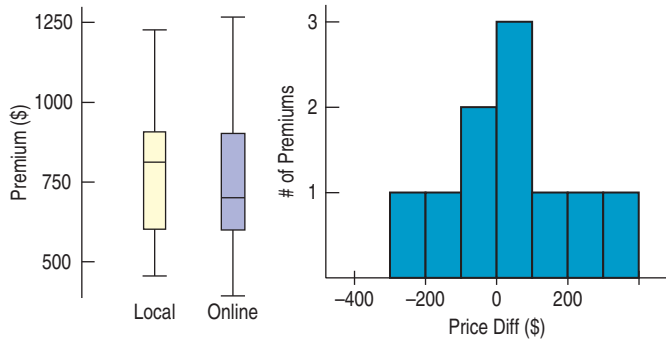
At first, the insurance agent wondered whether there was some kind of mistake in this output. He thought the Pythagorean Theorem of Statistics should work for finding the standard deviation of the price differences—in other words, that  $SD(Local - Online) = \sqrt{SD^2(Local) + SD^2(Online)}$ . But when he checked, he found that  $\sqrt{(229.281)^2 + (256.267)^2} = 343.864$ , not 175.663 as given by the software. Tell him where his mistake is.

- T** 10. **Windy, part I.** To select the site for an electricity-generating wind turbine, wind speeds were recorded at several potential sites every 6 hours for a year. Two sites not far from each other looked good. Each had a mean wind speed high enough to qualify, but we should choose the site with a higher average daily wind speed. Because the sites are near each other and the wind speeds were recorded at the same times, we should view the speeds as paired. Here are the summaries of the speeds (in miles per hour):

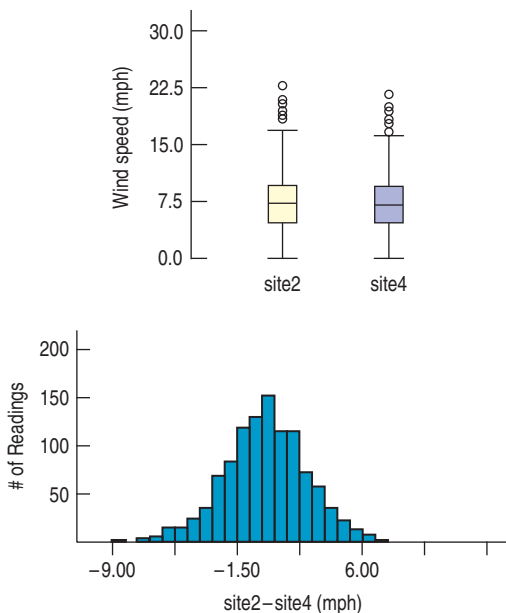
Variable	Count	Mean	StdDev
site2	1114	7.452	3.586
site4	1114	7.248	3.421
site2 - site4	1114	0.204	2.551

Is there a mistake in this output? Why doesn't the Pythagorean Theorem of Statistics work here? In other words, shouldn't  $SD(site2 - site4) = \sqrt{SD^2(site2) + SD^2(site4)}$ ? But  $\sqrt{(3.586)^2 + (3.421)^2} = 4.956$ , not 2.551 as given by the software. Explain why this happened.

**11. Online insurance II.** In Exercise 9, we saw summary statistics for 10 drivers' car insurance premiums quoted by a local agent and an online company. Here are displays for each company's quotes and for the difference (*Local* - *Online*):



- a) Which of the summaries would help you decide whether the online company offers cheaper insurance? Why?
  - b) The standard deviation of *PriceDiff* is quite a bit smaller than the standard deviation of prices quoted by either the local or online companies. Discuss why.
  - c) Using the information you have, discuss the assumptions and conditions for inference with these data.
- T 12. Windy, part II.** In Exercise 10, we saw summary statistics for wind speeds at two sites near each other, both being considered as locations for an electricity-generating wind turbine. The data, recorded every 6 hours for a year, showed each of the sites had a mean wind speed high enough to qualify, but how can we tell which site is best? Here are some displays:



- a) The boxplots show outliers for each site, yet the histogram shows none. Discuss why.
- b) Which of the summaries would you use to select between these sites? Why?
- c) Using the information you have, discuss the assumptions and conditions for paired *t* inference for these data. (*Hint*: Think hard about the independence assumption in particular.)

**13. Online insurance 3.** Exercises 9 and 11 give summaries and displays for car insurance premiums quoted by a local agent and an online company. Test an appropriate hypothesis to see if there is evidence that drivers might save money by switching to the online company.

**T 14. Windy, part III.** Exercises 10 and 12 give summaries and displays for two potential sites for a wind turbine. Test an appropriate hypothesis to see if there is evidence that either of these sites has a higher average wind speed.

**T 15. Temperatures.** The table below gives the average high temperatures in January and July for several European cities. Write a 90% confidence interval for the mean temperature difference between summer and winter in Europe. Be sure to check conditions for inference, and clearly explain what your interval means.

City	Mean High Temperatures (°F)	
	Jan.	July
Vienna	34	75
Copenhagen	36	72
Paris	42	76
Berlin	35	74
Athens	54	90
Rome	54	88
Amsterdam	40	69
Madrid	47	87
London	44	73
Edinburgh	43	65
Moscow	21	76
Belgrade	37	84

**T 16. Marathons 2006.** The table on the next page shows the winning times (in minutes) for men and women in the New York City Marathon between 1978 and 2006. Assuming that performances in the Big Apple resemble performances elsewhere, we can think of these data as a sample of performance in marathon competitions. Create a 90% confidence interval for the mean difference in winning times for male and female marathon competitors. ([www.nycmarathon.org](http://www.nycmarathon.org))

Year	Men	Women	Year	Men	Women
1978	132.2	152.5	1993	130.1	146.4
1979	131.7	147.6	1994	131.4	147.6
1980	129.7	145.7	1995	131.0	148.1
1981	128.2	145.5	1996	129.9	148.3
1982	129.5	147.2	1997	128.2	148.7
1983	129.0	147.0	1998	128.8	145.3
1984	134.9	149.5	1999	129.2	145.1
1985	131.6	148.6	2000	130.2	145.8
1986	131.1	148.1	2001	127.7	144.4
1987	131.0	150.3	2002	128.1	145.9
1988	128.3	148.1	2003	130.5	142.5
1989	128.0	145.5	2004	129.5	143.2
1990	132.7	150.8	2005	129.5	144.7
1991	129.5	147.5	2006	130.0	145.1
1992	129.5	144.7			

- T 17. Push-ups.** Every year the students at Gossett High School take a physical fitness test during their gym classes. One component of the test asks them to do as many push-ups as they can. Results for one class are shown below, separately for boys and girls. Assuming that students at Gossett are assigned to gym classes at random, create a 90% confidence interval for how many more push-ups boys can do than girls, on average, at that high school.

Boys	17	27	31	17	25	32	28	23	25	16	11	34
Girls	24	7	14	16	2	15	19	25	10	27	31	8

- T 18. Brain waves.** An experiment was performed to see whether sensory deprivation over an extended period of time has any effect on the alpha-wave patterns produced by the brain. To determine this, 20 subjects, inmates in a Canadian prison, were randomly split into two groups. Members of one group were placed in solitary confinement. Those in the other group were allowed to remain in their own cells. Seven days later, alpha-wave frequencies were measured for all subjects, as shown in the following table. (P. Gendreau et al., "Changes in EEG Alpha Frequency and Evoked Response Latency During Solitary Confinement," *Journal of Abnormal Psychology* 79 [1972]: 54–59)

Nonconfined	Confined
10.7	9.6
10.7	10.4
10.4	9.7
10.9	10.3
10.5	9.2
10.3	9.3
9.6	9.9
11.1	9.5
11.2	9.0
10.4	10.9

- What are the null and alternative hypotheses? Be sure to define all the terms and symbols you use.
- Are the assumptions necessary for inference met?
- Perform the appropriate test, indicating the formula you used, the calculated value of the test statistic, the df, and the P-value.
- State your conclusion.

- T 19. Job satisfaction.** (When you first read about this exercise break plan in Chapter 24, you did not have an inference method that would work. Try again now.) A company institutes an exercise break for its workers to see if it will improve job satisfaction, as measured by a questionnaire that assesses workers' satisfaction. Scores for 10 randomly selected workers before and after the implementation of the exercise program are shown in the table below.
- Identify the procedure you would use to assess the effectiveness of the exercise program, and check to see if the conditions allow the use of that procedure.
  - Test an appropriate hypothesis and state your conclusion.
  - If your conclusion turns out to be incorrect, what kind of error did you commit?

Worker Number	Job Satisfaction Index	
	Before	After
1	34	33
2	28	36
3	29	50
4	45	41
5	26	37
6	27	41
7	24	39
8	15	21
9	15	20
10	27	37

- T 20. Summer school.** (When you first read about the summer school issue in Chapter 24 you did not have an inference method that would work. Try again now.) Having done poorly on their Math final exams in June, six students repeat the course in summer school and take another exam in August.

June	54	49	68	66	62	62
Aug	50	65	74	64	68	72

- If we consider these students to be representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile?
  - This conclusion, of course, may be incorrect. If so, which type of error was made?
- T 21. Yogurt.** Is there a significant difference in calories between servings of strawberry and vanilla yogurt? Based on the data shown in the table, test an appropriate

hypothesis and state your conclusion. Don't forget to check assumptions and conditions!

	Calories per Serving	
	Strawberry	Vanilla
America's Choice	210	200
Breyer's Lowfat	220	220
Columbo	220	180
Dannon Light 'n Fit	120	120
Dannon Lowfat	210	230
Dannon la Crème	140	140
Great Value	180	80
La Yogurt	170	160
Mountain High	200	170
Stonyfield Farm	100	120
Yoplait Custard	190	190
Yoplait Light	100	100

- T 22. Gasoline.** Many drivers of cars that can run on regular gas actually buy premium in the belief that they will get better gas mileage. To test that belief, we use 10 cars from a company fleet in which all the cars run on regular gas. Each car is filled first with either regular or premium gasoline, decided by a coin toss, and the mileage for that tankful is recorded. Then the mileage is recorded again for the same cars for a tankful of the other kind of gasoline. We don't let the drivers know about this experiment. Here are the results (miles per gallon):

Car #	1	2	3	4	5	6	7	8	9	10
Regular	16	20	21	22	23	22	27	25	27	28
Premium	19	22	24	24	25	25	26	26	28	32

- Is there evidence that cars get significantly better fuel economy with premium gasoline?
  - How big might that difference be? Check a 90% confidence interval.
  - Even if the difference is significant, why might the company choose to stick with regular gasoline?
  - Suppose you had done a "bad thing." (We're sure you didn't.) Suppose you had mistakenly treated these data as two independent samples instead of matched pairs. What would the significance test have found? Carefully explain why the results are so different.
- T 23. Braking test.** A tire manufacturer tested the braking performance of one of its tire models on a test track. The company tried the tires on 10 different cars, recording the stopping distance for each car on both wet and dry pavement. Results are shown in the table.

Car #	Stopping Distance (ft)	
	Dry Pavement	Wet Pavement
1	150	201
2	147	220
3	136	192
4	134	146
5	130	182
6	134	173
7	134	202
8	128	180
9	136	192
10	158	206

- Write a 95% confidence interval for the mean dry pavement stopping distance. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.
  - Write a 95% confidence interval for the mean increase in stopping distance on wet pavement. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.
- T 24. Braking test 2.** For another test of the tires in Exercise 23, a car made repeated stops from 60 miles per hour. The test was run on both dry and wet pavement, with results as shown in the table. (Note that actual *braking distance*, which takes into account the driver's reaction time, is much longer, typically nearly 300 feet at 60 mph!)
- Write a 95% confidence interval for the mean dry pavement stopping distance. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.
  - Write a 95% confidence interval for the mean increase in stopping distance on wet pavement. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.

Stopping Distance (ft)	
Dry Pavement	Wet Pavement
145	211
152	191
141	220
143	207
131	198
148	208
126	206
140	177
135	183
133	223



- T 25. Tuition 2006.** How much more do public colleges and universities charge out-of-state students for tuition per semester? A random sample of 19 public colleges and universities listed at [www.collegeboard.com](http://www.collegeboard.com) yielded the following data. Tuition figures per semester are rounded to the nearest hundred dollars.

Institution	Resident	Nonresident
Univ of Akron (OH)	4200	8800
Athens State (AL)	1900	3600
Ball State (IN)	3400	8600
Bloomsburg U (PA)	3200	7000
UC Irvine (CA)	3400	12700
Central State (OH)	2600	5700
Clarion U (PA)	3300	5900
Dakota State	2900	3400
Fairmont State (WV)	2200	4600
Johnson State (VT)	3400	7300
Lock Haven U (PA)	3200	6000
New College of Florida	1600	8300
Oakland U (MI)	3300	7700
U Pittsburgh	6100	10700
Savannah State (GA)	1600	5400
SE Louisiana	1700	4400
W Liberty State (WV)	2000	4800
W Texas College	800	1000
Worcester State (MA)	2800	5800

- Create a 90% confidence interval for the mean difference in cost. Be sure to justify your procedure.
- Interpret your interval in context.
- A national magazine claims that public institutions charge state residents an average of \$3500 less than out-of-staters for tuition each semester. What does your confidence interval indicate about this assertion?

- T 26. Sex sells, part II.** In Exercise 3 you considered the question of whether sexual images in ads affected people's abilities to remember the item being advertised. To investigate, a group of Statistics students cut ads out of magazines. They were careful to find two ads for each of 10 similar items, one with a sexual image and one without. They arranged the ads in random order and had 39 subjects look at them for one minute. Then they asked the subjects to list as many of the products as they could remember. Their data are shown in the table. Is there evidence that the sexual images mattered?

Subject Number	Ads Remembered		Subject Number	Ads Remembered	
	Sexual Image	No Sex		Sexual Image	No Sex
1	2	2	21	2	3
2	6	7	22	4	2
3	3	1	23	3	3
4	6	5	24	5	3
5	1	0	25	4	5
6	3	3	26	2	4
7	3	5	27	2	2
8	7	4	28	2	4
9	3	7	29	7	6
10	5	4	30	6	7
11	1	3	31	4	3
12	3	2	32	4	5
13	6	3	33	3	0
14	7	4	34	4	3
15	3	2	35	2	3
16	7	4	36	3	3
17	4	4	37	5	5
18	1	3	38	3	4
19	5	5	39	4	3
20	2	2			

- T 27. Strikes.** Advertisements for an instructional video claim that the techniques will improve the ability of Little League pitchers to throw strikes and that, after undergoing the training, players will be able to throw strikes on at least 60% of their pitches. To test this claim, we have 20 Little Leaguers throw 50 pitches each, and we record the number of strikes. After the players participate in the training program, we repeat the test. The table shows the number of strikes each player threw before and after the training.
- Is there evidence that after training players can throw strikes more than 60% of the time?
  - Is there evidence that the training is effective in improving a player's ability to throw strikes?

Number of Strikes (out of 50)		Number of Strikes (out of 50)	
Before	After	Before	After
28	35	33	33
29	36	33	35
30	32	34	32
32	28	34	30
32	30	34	33
32	31	35	34
32	32	36	37
32	34	36	33
32	35	37	35
33	36	37	32

- T 28. Freshman 15, revisited.** In Exercise 4 you thought about how to design a study to see if it's true that students tend to gain weight during their first year in college. Well, Cornell Professor of Nutrition David Levitsky did just that. He recruited students from two large sections of an introductory health course. Although they were volunteers, they appeared to match the rest of the freshman class in terms of demographic variables such as sex and ethnicity. The students were weighed during the first week of the semester, then again 12 weeks later. Based on Professor Levitsky's data, estimate the mean weight gain in first-semester freshmen and comment on the "freshman 15." (Weights are in pounds.)

Subject Number	Initial Weight	Terminal Weight	Subject Number	Initial Weight	Terminal Weight
1	171	168	35	148	150
2	110	111	36	164	165
3	134	136	37	137	138
4	115	119	38	198	201
5	150	155	39	122	124
6	104	106	40	146	146
7	142	148	41	150	151
8	120	124	42	187	192
9	144	148	43	94	96
10	156	154	44	105	105
11	114	114	45	127	130
12	121	123	46	142	144
13	122	126	47	140	143
14	120	115	48	107	107
15	115	118	49	104	105
16	110	113	50	111	112
17	142	146	51	160	162
18	127	127	52	134	134
19	102	105	53	151	151
20	125	125	54	127	130
21	157	158	55	106	108
22	119	126	56	185	188
23	113	114	57	125	128
24	120	128	58	125	126
25	135	139	59	155	158
26	148	150	60	118	120
27	110	112	61	149	150
28	160	163	62	149	149
29	220	224	63	122	121
30	132	133	64	155	158
31	145	147	65	160	161
32	141	141	66	115	119
33	158	160	67	167	170
34	135	134	68	131	131



### JUST CHECKING Answers

- These are independent groups sampled at random, so use a two-sample  $t$  confidence interval to estimate the size of the difference.
- There is only one sample. Use a one-sample  $t$ -interval.
- A brother and sister from the same family represent a matched pair. The question calls for a paired  $t$ -test.
- A before-and-after study calls for paired  $t$ -methods. To estimate the loss, find a confidence interval for the before-after differences.
  - The two treatment groups were assigned randomly, so they are independent. Use a two-sample  $t$ -test to assess whether the mean weight losses differ.
- Sometimes it just isn't clear. Most likely, couples would discuss the band or even decide to go to the club because they both like a particular band. If we think that's likely, then these data are paired. But maybe not. If we asked them their opinions of, say, the decor or furnishings at the club, the fact that they were couples might not affect the independence of their answers.

## REVIEW OF PART VI

## Learning About the World

## Quick Review

We continue to explore how to answer questions about the statistics we get from samples and experiments. In this part, those questions have been about means—means of one sample, two independent samples, or matched pairs. Here's a brief summary of the key concepts and skills:

- ▶ A confidence interval uses a sample statistic to estimate a range of possible values for a parameter of interest.
- ▶ A hypothesis test proposes a model, then examines the plausibility of that model by seeing how surprising our observed data would be if the model were true.
- ▶ Statistical inference procedures for proportions are based on the Central Limit Theorem. We can make inferences about a single proportion or the difference of two proportions using Normal models.
- ▶ Statistical inference procedures for means are also based on the Central Limit Theorem, but we don't usually know the population standard deviation. Student's  $t$ -models take into account the additional uncertainty of independently estimating the standard deviation.
  - We can make inferences about one mean, the difference of two independent means, or the mean of paired differences using  $t$ -models.
  - No inference procedure is valid unless the underlying assumptions are true. Always check the conditions before proceeding.

- Because  $t$ -models assume that samples are drawn from Normal populations, data in the sample should appear to be nearly Normal. Skewness and outliers are particularly problematic, especially for small samples.
- When there are two variables, you must think carefully about how the data were collected. You may use two-sample  $t$  procedures only if the groups are independent.
- Unless there is some obvious reason to suspect that two independent populations have the same standard deviation, you should not pool the variances. It is never wrong to use unpooled  $t$  procedures.
- If the two groups are somehow paired, the data are *not* from independent groups. You must use matched-pairs  $t$  procedures.

Now for some opportunities to review these concepts. Be careful. You have a lot of thinking to do. These review exercises mix questions about proportions and means. You have to determine which of our inference procedures is appropriate in each situation. Then you have to check the proper assumptions and conditions. Keeping track of those can be difficult, so first we summarize the many procedures with their corresponding assumptions and conditions on the next page. Look them over carefully . . . then, on to the Exercises!

Quick Guide to Inference

Think			Show				Tell?
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Proportions	One sample	1-Proportion z-Interval	z	p	$\hat{p}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$	19
		1-Proportion z-Test				$\sqrt{\frac{p_0q_0}{n}}$	20, 21
	Two independent groups	2-Proportion z-Interval	z	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	22
		2-Proportion z-Test				$\sqrt{\frac{\hat{p}\hat{q}}{n_1} + \frac{\hat{p}\hat{q}}{n_2}}, \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$	22
Means	One sample	t-Interval t-Test	t df = n - 1	$\mu$	$\bar{y}$	$\frac{s}{\sqrt{n}}$	23
	Two independent groups	2-Sample t-Test 2-Sample t-Interval	t df from technology	$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	24
	Matched pairs	Paired t-Test Paired t-Interval	t df = n - 1	$\mu_d$	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$	25

Assumptions for Inference

And the Conditions That Support or Override Them

Proportions (z)

• One sample

1. Individuals are independent.
2. Sample is sufficiently large.

• Two groups

1. Groups are independent.
2. Data in each group are independent.
3. Both groups are sufficiently large.

1. SRS and  $n < 10\%$  of the population.
2. Successes and failures each  $\geq 10$ .

1. (Think about how the data were collected.)
2. Both are SRSs and  $n < 10\%$  of populations OR random allocation.
3. Successes and failures each  $\geq 10$  for both groups.

Means (t)

• One sample (df = n - 1)

1. Individuals are independent.
2. Population has a Normal model.

• Matched pairs (df = n - 1)

1. Data are matched.
2. Individuals are independent.
3. Population of differences is Normal.

• Two independent groups (df from technology)

1. Groups are independent.
2. Data in each group are independent.
3. Both populations are Normal.

1. SRS and  $n < 10\%$  of the population.
2. Histogram is unimodal and symmetric.\*

1. (Think about the design.)
2. SRS and  $n < 10\%$  OR random allocation.
3. Histogram of differences is unimodal and symmetric.\*

1. (Think about the design.)
2. SRSs and  $n < 10\%$  OR random allocation.
3. Both histograms are unimodal and symmetric.\*

(\*less critical as n increases)

## REVIEW EXERCISES

1. **Crawling.** A study published in 1993 found that babies born at different times of the year may develop the ability to crawl at different ages! The author of the study suggested that these differences may be related to the temperature at the time the infant is 6 months old. (Benson and Janette, *Infant Behavior and Development* [1993])

- The study found that 32 babies born in January crawled at an average age of 29.84 weeks, with a standard deviation of 7.08 weeks. Among 21 July babies, crawling ages averaged 33.64 weeks, with a standard deviation of 6.91 weeks. Is this difference significant?
- For 26 babies born in April the mean and standard deviation were 31.84 and 6.21 weeks, while for 44 October babies the mean and standard deviation of crawling ages were 33.35 and 7.29 weeks. Is this difference significant?
- Are these results consistent with the researcher's conjecture?

**T** 2. **Mazes and smells.** Can pleasant smells improve learning? Researchers timed 21 subjects as they tried to complete paper-and-pencil mazes. Each subject attempted a maze both with and without the presence of a floral aroma. Subjects were randomized with respect to whether they did the scented trial first or second. Is there any evidence that the floral scent improved the subjects' ability to complete the mazes? (A. R. Hirsch and L. H. Johnston, "Odors and Learning." Chicago: Smell and Taste Treatment and Research Foundation)

Time to Complete the Maze (sec)	
Unscented	Scented
25.7	30.2
41.9	56.7
51.9	42.4
32.2	34.4
64.7	44.8
31.4	42.9
40.1	42.7
43.2	24.8
33.9	25.1
40.4	59.2
58.0	42.2
61.5	48.4
44.6	32.0
35.3	48.1
37.2	33.7
39.4	42.6
77.4	54.9
52.8	64.5
63.6	43.1
56.6	52.8
58.9	44.3

3. **Women.** The U.S. Census Bureau reports that 26% of all U.S. businesses are owned by women. A Colorado consulting firm surveys a random sample of 410 businesses in the Denver area and finds that 115 of them have women owners. Should the firm conclude that its area is unusual? Test an appropriate hypothesis and state your conclusion.

**T** 4. **Drugs.** In a full-page ad that ran in many U.S. newspapers in August 2002, a Canadian discount pharmacy listed costs of drugs that could be ordered from a Web site in Canada. The table compares prices (in US\$) for commonly prescribed drugs.

Drug Name	Cost per 100 Pills		Percent savings
	United States	Canada	
Cardizem	131	83	37
Celebrex	136	72	47
Cipro	374	219	41
Pravachol	370	166	55
Premarin	61	17	72
Prevacid	252	214	15
Prozac	263	112	57
Tamoxifen	349	50	86
Vioxx	243	134	45
Zantac	166	42	75
Zocor	365	200	45
Zoloft	216	105	51

- Give a 95% confidence interval for the average savings in dollars.
- Give a 95% confidence interval for the average savings in percent.
- Which analysis is more appropriate? Why?
- In small print the newspaper ad says, "Complete list of all 1500 drugs available on request." How does this comment affect your conclusions above?

**T** 5. **Pottery.** Archaeologists can use the chemical composition of clay found in pottery artifacts to determine whether different sites were populated by the same ancient people. They collected five samples of Romano-British pottery from each of two sites in Great Britain and measured the percentage of aluminum oxide in each. Based on these data, do you think the same people used these two kiln sites? Base your conclusion on a 95% confidence interval for the difference in aluminum oxide content of pottery made at the sites. (A. Tubb, A. J. Parker, and G. Nickless, "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry." *Archaeometry*, 22[1980]:153–171)

Ashley Rails	19.1	14.8	16.7	18.3	17.7
New Forest	20.8	18.0	18.0	15.8	18.3

6. **Streams.** Researchers in the Adirondack Mountains collect data on a random sample of streams each year. One of the variables recorded is the substrate of the streams—the type of soil and rock over which they flow. The researchers found that 69 of the 172 sampled streams had a substrate of shale. Construct a 95% confidence interval for the proportion of Adirondack streams with a shale substrate. Clearly interpret your interval in context.

7. **Gehrig.** Ever since Lou Gehrig developed amyotrophic lateral sclerosis (ALS), this deadly condition has been commonly known as Lou Gehrig's disease. Some believe that ALS is more likely to strike athletes or the very fit. Columbia University neurologist Lewis P. Rowland recorded personal histories of 431 patients he examined between 1992 and 2002. He diagnosed 280 as having ALS; 38% of them had been varsity athletes. The other 151 had other neurological disorders, and only 26% of them had been varsity athletes. (*Science News*, Sept. 28 [2002])

- Is there evidence that ALS is more common among athletes?
- What kind of study is this? How does that affect the inference you made in part a?

**T** 8. **Teen drinking.** A study of the health behavior of school-aged children asked a sample of 15-year-olds in several different countries if they had been drunk at least twice. The results are shown in the table, by gender. Give a 95% confidence interval for the difference in the rates for males and females. Be sure to check the assumptions that support your chosen procedure, and explain what your interval means. (*Health and Health Behavior Among Young People*. Copenhagen: World Health Organization, 2000)

Country	Percent of 15-Year-Olds Drunk at Least Twice	
	Female	Male
Denmark	63	71
Wales	63	72
Greenland	59	58
England	62	51
Finland	58	52
Scotland	56	53
No. Ireland	44	53
Slovakia	31	49
Austria	36	49
Canada	42	42
Sweden	40	40
Norway	41	37
Ireland	29	42
Germany	31	36
Latvia	23	47
Estonia	23	44
Hungary	22	43
Poland	21	39
USA	29	34
Czech Rep.	22	36
Belgium	22	36
Russia	25	32
Lithuania	20	32
France	20	29
Greece	21	24
Switzerland	16	25
Israel	10	18

9. **Babies.** The National Perinatal Statistics Unit of the Sydney Children's Hospital reports that the mean birth weight of all babies born in Australia in 1999 was 3361 grams—about 7.41 pounds. A Missouri hospital reports that the average weight of 112 babies born there last year was 7.68 pounds, with a standard deviation of 1.31 pounds. If we believe the Missouri babies fairly represent American newborns, is there any evidence that U.S. babies and Australian babies do not weigh the same amount at birth?

- Petitions.** To get a voter initiative on a state ballot, petitions that contain at least 250,000 valid voter signatures must be filed with the Elections Commission. The board then has 60 days to certify the petitions. A group wanting to create a statewide system of universal health insurance has just filed petitions with a total of 304,266 signatures. As a first step in the process, the Board selects an SRS of 2000 signatures and checks them against local voter lists. Only 1772 of them turn out to be valid.
  - What percent of the sample signatures were valid?
  - What percent of the petition signatures submitted must be valid in order to have the initiative certified by the Elections Commission?
  - What will happen if the Elections Commission commits a Type I error?
  - What will happen if the Elections Commission commits a Type II error?
  - Does the sample provide evidence in support of certification? Explain.
  - What could the Elections Commission do to increase the power of the test?

11. **Feeding fish.** In the midwestern United States, a large aquaculture industry raises largemouth bass. Researchers wanted to know whether the fish would grow better if fed a natural diet of fathead minnows or an artificial diet of food pellets. They stocked six ponds with bass fingerlings weighing about 8 grams. For one year, the fish in three of the ponds were fed minnows, and the others were fed the commercially prepared pellets. The fish were then harvested, weighed, and measured. The bass fed a natural food source had a higher average length (19.6 cm) and weight (95.9 g) than those fed the commercial fish food (17.3 cm and 72.0 g, respectively). The researchers reported P-values for differences in both measurements to be less than 0.001.

- Explain to someone who has not studied Statistics what the P-values mean here.
- What advice should the researchers give the people who raise largemouth bass?
- If that advice turns out to be incorrect, what type of error occurred?

**T** 12. **Risk.** A study of auto safety determined the number of driver deaths per million vehicle sales, classified by type of vehicle. The data on the next page are for 6 midsize

models and 6 SUVs. Wondering if there is evidence that drivers of SUVs are safer, we hope to create a 95% confidence interval for the difference in driver death rates for the two types of vehicles. Are these data appropriate for this inference? Explain. (Ross and Wenzel, *An Analysis of Traffic Deaths by Vehicle Type and Model*, March 2002)

<b>Midsized</b>	47	54	64	76	88	97
<b>SUV</b>	55	60	62	76	91	109

13. **Age.** In a study of how depression may affect one's ability to survive a heart attack, the researchers reported the ages of the two groups they examined. The mean age of 2397 patients without cardiac disease was 69.8 years ( $SD = 8.7$  years), while for the 450 patients with cardiac disease, the mean and standard deviation of the ages were 74.0 and 7.9, respectively.
- Create a 95% confidence interval for the difference in mean ages of the two groups.
  - How might an age difference confound these research findings about the relationship between depression and ability to survive a heart attack?
14. **Smoking.** In the depression and heart attack research described in Exercise 13, 32% of the diseased group were smokers, compared with only 23.7% of those free of heart disease.
- Create a 95% confidence interval for the difference in the proportions of smokers in the two groups.
  - Is this evidence that the two groups in the study were different? Explain.
  - Could this be a problem in analyzing the results of the study? Explain.
15. **Computer use.** A Gallup telephone poll of 1240 teens conducted in 2001 found that boys were more likely than girls to play computer games, by a margin of 77% to 65%. Equal numbers of boys and girls were surveyed.
- What kind of sampling design was used?
  - Give a 95% confidence interval for the difference in game playing by gender.
  - Does your confidence interval suggest that among all teens a higher percentage of boys than girls play computer games?
16. **Recruiting.** In September 2002, CNN reported on a method of grad student recruiting by the Haas School of Business at U.C.-Berkeley. The school notifies applicants by formal letter that they have been admitted, and also e-mails the accepted students a link to a Web site that greets them with personalized balloons, cheering, and applause. The director of admissions says this extra effort at recruiting has really worked well. The school accepts 500 applicants each year, and the percentage that actually choose to enroll at Berkeley increased from 52% the year before the Web greeting to 54% this year.
- Create a 95% confidence interval for the change in enrollment rates.
- b) Based on your confidence interval, are you convinced that this new form of recruiting has been effective? Explain.
- T 17. Hearing.** Fitting someone for a hearing aid requires assessing the patient's hearing ability. In one method of assessment, the patient listens to a tape of 50 English words. The tape is played at low volume, and the patient is asked to repeat the words. The patient's hearing ability score is the number of words perceived correctly. Four tapes of equivalent difficulty are available so that each ear can be tested with more than one hearing aid. These lists were created to be equally difficult to perceive in silence, but hearing aids must work in the presence of background noise. Researchers had 24 subjects with normal hearing compare two of the tapes when a background noise was present, with the order of the tapes randomized. Is it reasonable to assume that the two lists are still equivalent for purposes of the hearing test when there is background noise? Base your decision on a confidence interval for the mean difference in the number of words people might misunderstand. (Faith Loven, *A Study of the Interlist Equivalency of the CID W-22 Word List Presented in Quiet and in Noise*. University of Iowa [1981])
18. **Cesareans.** Some people fear that differences in insurance coverage can affect healthcare decisions. A survey of several randomly selected hospitals found that 16.6% of 223 recent births in Vermont involved cesarean deliveries, compared to 18.8% of 186 births in New Hampshire. Is this evidence that the rate of cesarean births in the two states is different?
- T 19. Newspapers.** Who reads the newspaper more, men or women? Eurostat, an agency of the European Union (EU), conducts surveys on several aspects of daily life in EU countries. Recently, the agency asked samples of 1000 respondents in each of 14 European countries whether they read the newspaper on a daily basis. The table on the next page shows the data.

Subject	List A	List B
1	24	26
2	32	24
3	20	22
4	14	18
5	32	24
6	22	30
7	20	22
8	26	28
9	26	30
10	38	16
11	30	18
12	16	34
13	36	32
14	32	34
15	38	32
16	14	18
17	26	20
18	14	20
19	38	40
20	20	26
21	14	14
22	18	14
23	22	30
24	34	42

% Reading a Newspaper Daily		
Country	Men	Women
Belgium	56.3	45.5
Denmark	76.8	70.3
Germany	79.9	76.8
Greece	22.5	17.2
Spain	46.2	24.8
Ireland	58.0	54.0
Italy	50.2	29.8
Luxembourg	71.0	67.0
Netherlands	71.3	63.0
Austria	78.2	74.1
Portugal	58.3	24.1
Finland	93.0	90.0
Sweden	89.0	88.0
UK	32.6	30.4

- a) Examine the differences in the percentages for each country. Which of these countries seem to be outliers? What do they have in common?
- b) After eliminating the outliers, is there evidence that in Europe men are more likely than women to read the newspaper?
- T** 20. **Meals.** A college student is on a “meal program.” His budget allows him to spend an average of \$10 per day for the semester. He keeps track of his daily food expenses for 2 weeks; the data are given in the table. Is there strong evidence that he will overspend his food allowance? Explain.

Date	Cost (\$)	Date	Cost (\$)
7/29	15.20	8/5	8.55
7/30	23.20	8/6	20.05
7/31	3.20	8/7	14.95
8/1	9.80	8/8	23.45
8/2	19.53	8/9	6.75
8/3	6.25	8/10	0
8/4	0	8/11	9.01

21. **Wall Street.** In September of 2000, the Harris Poll organization asked 1002 randomly sampled American adults whether they agreed or disagreed with the following statement:

*Most people on Wall Street would be willing to break the law if they believed they could make a lot of money and get away with it.*

Of those asked, 60% said they agreed with this statement. We know that if we could ask the entire population of American adults, we would not find that exactly 60% think that Wall Street workers would be willing to break the law to make money. Construct a 95% confidence interval for the true percentage of American adults who agree with the statement.

22. **Teach for America.** Several programs attempt to address the shortage of qualified teachers by placing uncertified instructors in schools with acute needs—often in inner cities. A 1999–2000 study compared students taught by certified teachers to others taught by uncertified teachers in the same schools. Reading scores of the students of certified teachers averaged 35.62 points with standard deviation 9.31. The scores of students instructed by uncertified teachers had mean 32.48 points with standard deviation 9.43 points on the same test. There were 44 students in each group. The appropriate  $t$  procedure has 86 degrees of freedom. Is there evidence of lower scores with uncertified teachers? Discuss. (*The Effectiveness of “Teach for America” and Other Under-certified Teachers on Student Academic Achievement: A Case of Harmful Public Policy.* Education Policy Analysis Archives [2002])

- T** 23. **Legionnaires’ disease.** In 1974, the Bellevue-Stratford Hotel in Philadelphia was the scene of an outbreak of what later became known as legionnaires’ disease. The cause of the disease was finally discovered to be bacteria that thrived in the air-conditioning units of the hotel. Owners of the Rip Van Winkle Motel, hearing about the Bellevue-Stratford, replaced their air-conditioning system. The following data are the bacteria counts in the air of eight rooms, before and after a new air-conditioning system was installed (measured in colonies per cubic foot of air). Has the new system succeeded in lowering the bacterial count? Base your analysis on a confidence interval. Be sure to list all your assumptions, methods, and conclusions.

Room Number	Before	After
121	11.8	10.1
163	8.2	7.2
125	7.1	3.8
264	14	12
233	10.8	8.3
218	10.1	10.5
324	14.6	12.1
325	14	13.7

24. **Teach for America, Part II.** The study described in Exercise 22 also looked at scores in mathematics and language. Here are software outputs for the appropriate tests. Explain what they show.

**Mathematics**

T-TEST OF  $\mu(1) - \mu(2) = 0$   
 $\mu(\text{Cer } t) - \mu(\text{NoCer } t) = 4.53 \quad t(86) = 2.95 \quad p = 0.002$

**Language**

T-TEST OF  $\mu(1) - \mu(2) = 0$   
 $\mu(\text{Cer } t) - \mu(\text{NoCer } t) = 2.13 \quad t(84) = 1.71 \quad p = 0.045$

25. **Bipolar kids.** The June 2002 *American Journal of Psychiatry* reported that researchers used medication and psychotherapy to treat children aged 7 to 16 who exhibit bipolar symptoms. After 2 years, symptoms had cleared up in only 26 of the 89 children involved in the study.



- a) Write a 95% confidence interval; interpret it in context.  
 b) If researchers subsequently hope to produce an estimate of treatment effectiveness for bipolar disorder that has a margin of error of only 6%, how many patients should they study?

**T 26. Online testing.** The Educational Testing Service is now administering several of its standardized tests online—the CLEP and GMAT exams, for example. Since taking a test on a computer is different from taking a test with pencil and paper, one wonders if the scores will be the same. To investigate this question, researchers created two versions of an SAT-type test and got 20 volunteers to participate in an experiment. Each volunteer took both versions of the test, one with pencil and paper and the other online. Subjects were randomized with respect to the order in which they sat for the tests (online/paper) and which form they took (Test A, Test B) in which environment. The scores (out of a possible 20) are summarized in the table.

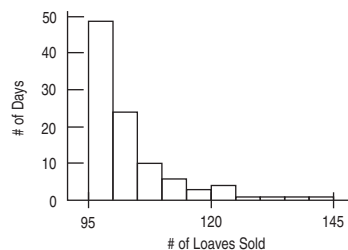
Subject	Paper	Online
	Test A	Test B
1	14	13
2	10	13
3	16	8
4	15	14
5	17	16
6	14	11
7	9	12
8	12	12
9	16	16
10	7	14
	Test B	Test A
11	8	13
12	11	13
13	15	17
14	11	13
15	13	14
16	9	9
17	15	9
18	14	15
19	16	12
20	8	10

- a) Were the two forms (A/B) of the test equivalent in terms of difficulty? Test an appropriate hypothesis and state your conclusion.  
 b) Is there evidence that the testing environment (paper/online) matters? Test an appropriate hypothesis and state your conclusion.

**27. Bread.** Clarksburg Bakery is trying to predict how many loaves of bread to bake. In the last 100 days, the bakery has sold between 95 and 140 loaves per day. Here are a histogram and the summary statistics for the number of loaves sold for the last 100 days.

#### Summary of Sales

Mean	103
Median	100
SD	9.000
Min	95
Max	140
$Q_1$	97
$Q_3$	105.5



- a) Can you use these data to estimate the number of loaves sold on the busiest 10% of all days? Explain.

- b) Explain why you can use these data to construct a 95% confidence interval for the mean number of loaves sold per day.  
 c) Calculate a 95% confidence interval and carefully interpret what that confidence interval means.  
 d) If the bakery would have been satisfied with a confidence interval whose margin of error was twice as wide, how many days' data could they have used?  
 e) When the bakery opened, the owners estimated that they would sell an average of 100 loaves per day. Does your confidence interval provide strong evidence that this estimate was incorrect? Explain.

**T 28. Irises.** Can measurements of the petal length of flowers be of value when you need to determine the species of a certain flower? Here are the summary statistics from measurements of the petals of two species of irises. (R. A. Fisher, "The Use of Multiple Measurements in Axonomic Problems." *Annals of Eugenics* 7 [1936]:179–188)

	Species	
	Versicolor	Virginica
Count	50	50
Mean	55.52	43.22
Median	55.50	44.00
SD	5.519	5.362
Min	45	30
Max	69	56
Lower Quartile	51	40
Upper Quartile	59	47

- a) Make parallel boxplots of petal lengths for the two species.  
 b) Describe the differences seen in the boxplots.  
 c) Write a 95% confidence interval for the difference in petal length.  
 d) Explain what your interval means.  
 e) Based on your confidence interval, is there evidence of a difference in petal length? Explain.

**29. Insulin and diet.** A study published in the *Journal of the American Medical Association* examined people to see if they showed any signs of IRS (insulin resistance syndrome) involving major risk factors for Type 2 diabetes and heart disease. Among 102 subjects who consumed dairy products more than 35 times per week, 24 were identified with IRS. In comparison, IRS was identified in 85 of 190 individuals with the lowest dairy consumption, fewer than 10 times per week.

- a) Is this strong evidence that IRS risk is different in people who frequently consume dairy products than in those who do not?  
 b) Does this indicate that dairy consumption influences the development of IRS? Explain.

**30. Speeding.** A newspaper report in August 2002 raised the issue of racial bias in the issuance of speeding tickets. The following facts were noted:

- 16% of drivers registered in New Jersey are black.

- Of the 324 speeding tickets issued in one month on a 65-mph section of the New Jersey Turnpike, 25% went to black drivers.
- a) Is the percentage of speeding tickets issued to blacks unusually high compared to registrations?
  - b) Does this suggest that racial profiling may be present?
  - c) What other statistics would you like to know about this situation?

**T 31. Rainmakers?** In an experiment to determine whether seeding clouds with silver iodide increases rainfall, 52 clouds were randomly assigned to be seeded or not. The amount of rain they generated was then measured (in acre-feet). Create a 95% confidence interval for the average amount of additional rain created by seeding clouds. Explain what your interval means.

	Unseeded Clouds	Seeded Clouds
Count	26	26
Mean	164.588	441.985
Median	44.200	221.600
SD	278.426	650.787
IntQRRange	138.600	337.600
25 %ile	24.400	92.400
75 %ile	163	430

- 32. Fritos.** As a project for an introductory Statistics course, students checked 6 bags of Fritos marked with a net weight of 35.4 grams. They carefully weighed the contents of each bag, recording the following weights (in grams): 35.5, 35.3, 35.1, 36.4, 35.4, 35.5. Is there evidence that the mean weight of bags of Fritos is less than advertised?
- a) Write appropriate hypotheses.
  - b) Check the assumptions for inference.
  - c) Test your hypothesis using all 6 weights.
  - d) Retest your hypothesis with the one unusually high weight removed.
  - e) What would you conclude about the stated weight?

**T 33. Color or text?** In an experiment, 32 volunteer subjects are briefly shown seven cards, each displaying the name of a color printed in a different color (example: red, blue, and so on). The subject is asked to perform one of two tasks: memorize the order of the words or memorize the order of the colors. Researchers record the number of cards remembered correctly. Then the cards are shuffled and the subject is asked to perform the other task. The table displays the results for each subject. Is there any evidence that either the color or the written word dominates perception?

- a) What role does randomization play in this experiment?
- b) Test appropriate hypotheses and state your conclusion.

Subject	Color	Word	Subject	Color	Word
1	4	7	17	4	3
2	1	4	18	7	4
3	5	6	19	4	3
4	1	6	20	0	6
5	6	4	21	3	3
6	4	5	22	3	5
7	7	3	23	7	3
8	2	5	24	3	7
9	7	5	25	5	6
10	4	3	26	3	4
11	2	0	27	3	5
12	5	4	28	1	4
13	6	7	29	2	3
14	3	6	30	5	3
15	4	6	31	3	4
16	4	7	32	6	7

- 34. And it means?** Every statement about a confidence interval contains two parts: the level of confidence and the interval. Suppose that an insurance agent estimating the mean loss claimed by clients after home burglaries created the 95% confidence interval (\$1644, \$2391).
- a) What's the margin of error for this estimate?
  - b) Carefully explain what the interval means.
  - c) Carefully explain what the confidence level means.
- 35. Batteries.** We work for the "Watchdog for the Consumer" consumer advocacy group. We've been asked to look at a battery company that claims its batteries last an average of 100 hours under normal use. There have been several complaints that the batteries don't last that long, so we decide to test them. To do this, we select 16 batteries and run them until they die. They lasted a mean of 97 hours, with a standard deviation of 12 hours.
- a) One of the editors of our newsletter (who does not know statistics) says that 97 hours is a lot less than the advertised 100 hours, so we should reject the company's claim. Explain to him the problem with doing that.
  - b) What are the null and alternative hypotheses?
  - c) What assumptions must we make in order to proceed with inference?
  - d) At a 5% level of significance, what do you conclude?
  - e) Suppose that, in fact, the average life of the company's batteries is only 98 hours. Has an error been made in part d? If so, what kind?
- 36. Hamsters.** How large are hamster litters? Among 47 golden hamster litters recorded, there were an average of 7.72 baby hamsters, with a standard deviation of 2.5.
- a) Create and interpret a 90% confidence interval.
  - b) Would a 98% confidence interval have a larger or smaller margin of error? Explain.
  - c) How many litters must be used to estimate the average litter size to within 1 baby hamster with 95% confidence?



PART

VIII

# Inference When Variables Are Related

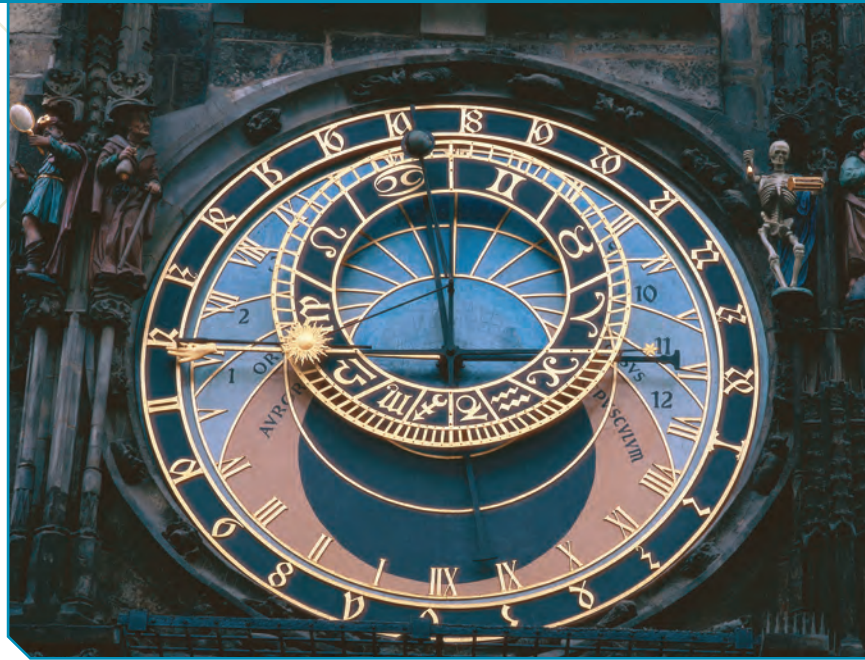
## Chapter 26

Comparing Counts

## Chapter 27

Inferences for Regression

# Comparing Counts



**WHO** Executives of Fortune 400 companies

**WHAT** Zodiac birth sign

**WHY** Maybe the researcher was a Gemini and naturally curious?

**A S** **Activity: Children at Risk.**  
See how a contingency table helps us understand the different risks to which an incident exposed children.

Does your zodiac sign predict how successful you will be later in life? *Fortune* magazine collected the zodiac signs of 256 heads of the largest 400 companies. The table shows the number of births for each sign.

We can see some variation in the number of births per sign, and there *are* more Pisces, but is that enough to claim that successful people are more likely to be born under some signs than others?

Births	Sign
23	Aries
20	Taurus
18	Gemini
23	Cancer
20	Leo
19	Virgo
18	Libra
21	Scorpio
19	Sagittarius
22	Capricorn
24	Aquarius
29	Pisces

Birth totals by sign for 256 Fortune 400 executives.

## Goodness-of-Fit

*“All creatures have their determined time for giving birth and carrying fetus, only a man is born all year long, not in determined time, one in the seventh month, the other in the eighth, and so on till the beginning of the eleventh month.”*

—Aristotle

If births were distributed uniformly across the year, we would expect about  $1/12$  of them to occur under each sign of the zodiac. That suggests  $256/12$ , or about 21.3 births per sign. How closely do the observed numbers of births per sign fit this simple “null” model?

A hypothesis test to address this question is called a test of “goodness-of-fit.” The name suggests a certain badness-of-grammar, but it is quite standard. After all, we are asking whether the model that births are uniformly distributed over the signs fits the data good, . . . er, well. Goodness-of-fit involves testing a hypothesis. We have specified a model for the distribution and want to know whether it fits. There is no single parameter to estimate, so a confidence interval wouldn’t make much sense.

If the question were about only one astrological sign (for example, “Are executives more likely to be Pisces?”<sup>1</sup>), we could use a one-proportion z-test and ask if

<sup>1</sup> A question actually asked us by someone who was undoubtedly a Pisces.

the true proportion of executives with that sign is equal to  $1/12$ . However, here we have 12 hypothesized proportions, one for each sign. We need a test that considers all of them together and gives an overall idea of whether the observed distribution differs from the hypothesized one.

## FOR EXAMPLE

### Finding expected counts

Birth month may not be related to success as a CEO, but what about on the ball field? It has been proposed by some researchers that children who are the older ones in their class at school naturally perform better in sports and that these children then get more coaching and encouragement. Could that make a difference in who makes it to the professional level in sports?

Baseball is a remarkable sport, in part because so much data are available. We have the birth dates of every one of the 16,804 players who ever played in a major league game. Since the effect we're suspecting may be due to relatively recent policies (and to keep the sample size moderate), we'll consider the birth months of the 1478 major league players born since 1975 and who have played through 2006. We can also look up the national demographic statistics to find what percentage of people were born in each month. Let's test whether the observed distribution of ballplayers' birth months shows just random fluctuations or whether it represents a real deviation from the national pattern.

**Question:** How can we find the expected counts?

There are 1478 players in this set of data. I'd expect 8% of them to have been born in January, and  $1478(0.08) = 118.24$ . I won't round off, because expected "counts" needn't be integers. Multiplying 1478 by each of the birth percentages gives the expected counts shown in the table.

Month	Ballplayer count	National birth %	Month	Ballplayer count	National birth %
1	137	8%	7	102	9%
2	121	7%	8	165	9%
3	116	8%	9	134	9%
4	121	8%	10	115	9%
5	126	8%	11	105	8%
6	114	8%	12	122	9%
			<b>Total</b>	1478	100%

Month	Expected	Month	Expected
1	118.24	7	133.02
2	103.46	8	133.02
3	118.24	9	133.02
4	118.24	10	133.02
5	118.24	11	118.24
6	118.24	12	133.02

## Assumptions and Conditions

These data are organized in tables as we saw in Chapter 3, and the assumptions and conditions reflect that. Rather than having an observation for each individual, we typically work with summary counts in categories. In our example, we don't see the birth signs of each of the 256 executives, only the totals for each sign.

**Counted Data Condition:** The data must be *counts* for the categories of a categorical variable. This might seem a simplistic, even silly condition. But many kinds of values can be assigned to categories, and it is unfortunately common to find the methods of this chapter applied incorrectly to proportions, percentages, or measurements just because they happen to be organized in a table. So check to be sure the values in each **cell** really are counts.

### INDEPENDENCE ASSUMPTION

**Independence Assumption:** The counts in the cells should be independent of each other. The easiest case is when the individuals who are counted in the cells are sampled independently from some population. That's what we'd like to have if we want to draw conclusions about that population. Randomness can arise in

other ways, though. For example, these Fortune 400 executives are not a random sample, but we might still think that their birth dates are randomly distributed throughout the year. If we want to generalize to a large population, we should check the Randomization Condition.

**Randomization Condition:** The individuals who have been counted should be a random sample from the population of interest.

## SAMPLE SIZE ASSUMPTION

We must have enough data for the methods to work. We usually check the following:

**Expected Cell Frequency Condition:** We should expect to see at least 5 individuals in each cell.

The Expected Cell Frequency Condition sounds like—and is, in fact, quite similar to—the condition that  $np$  and  $nq$  be at least 10 when we tested proportions. In our astrology example, assuming equal births in each month leads us to expect 21.3 births per month, so the condition is easily met here.

### FOR EXAMPLE

#### Checking assumptions and conditions

**Recap:** Are professional baseball players more likely to be born in some months than in others? We have observed and expected counts for the 1478 players born since 1975.

**Question:** Are the assumptions and conditions met for performing a goodness-of-fit test?

- ✓ **Counted Data Condition:** I have month-by-month counts of ballplayer births.
- ✓ **Independence Assumption:** These births were independent.
- ✓ **Randomization Condition:** Although they are not a random sample, we can take these players to be representative of players past and future.
- ✓ **Expected Cell Frequency Condition:** The expected counts range from 103.46 to 133.02, all much greater than 5.
- ✓ **10% Condition:** These 1478 players are less than 10% of the population of 16,804 players who have ever played (or will play) major league baseball.

It's okay to use these data for a goodness-of-fit test.

## Calculations

### NOTATION ALERT:

We compare the counts *observed* in each cell with the counts we *expect* to find. The usual notation uses  $O$ 's and  $E$ 's or abbreviations such as those we've used here. The method for finding the expected counts depends on the model.

We have observed a count in each category from the data, and have an expected count for each category from the hypothesized proportions. Are the differences just natural sampling variability, or are they so large that they indicate something important? It's natural to look at the *differences* between these observed and expected counts, denoted  $(Obs - Exp)$ . We'd like to think about the total of the differences, but just adding them won't work because some differences are positive, others negative. We've been in this predicament before—once when we looked at deviations from the mean and again when we dealt with residuals. In fact, these *are* residuals. They're just the differences between the observed data and the counts given by the (null) model. We handle these residuals in essentially the same way we did in regression: We square them. That gives us positive values and focuses attention on any cells with large differences from what we expected. Because the differences between observed and expected counts generally get larger the more data we have, we also need to get an idea of the *relative* sizes of the differences. To do that, we divide each squared difference by the expected count for that cell.

**NOTATION ALERT:**

The only use of the Greek letter  $\chi$  in Statistics is to represent this statistic and the associated sampling distribution. This is another violation of our “rule” that Greek letters represent population parameters. Here we are using a Greek letter simply to name a family of distribution models and a statistic.

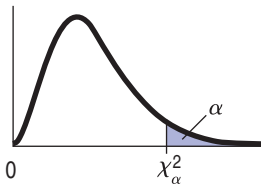
The test statistic, called the **chi-square** (or chi-squared) **statistic**, is found by adding up the sum of the squares of the deviations between the observed and expected counts divided by the expected counts:

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

The chi-square statistic is denoted  $\chi^2$ , where  $\chi$  is the Greek letter chi (pronounced “ky” as in “sky”). It refers to a family of sampling distribution models we have not seen before called (remarkably enough) the **chi-square models**.

This family of models, like the Student’s *t*-models, differ only in the number of degrees of freedom. The number of degrees of freedom for a goodness-of-fit test is  $n - 1$ . Here, however,  $n$  is *not* the sample size, but instead is the number of categories. For the zodiac example, we have 12 signs, so our  $\chi^2$  statistic has 11 degrees of freedom.

## One-Sided or Two-Sided?

**TI-*n*spire**

**The  $\chi^2$  Models.** See what a  $\chi^2$  model looks like, and watch it change as you change the degrees of freedom.

The chi-square statistic is used only for testing hypotheses, not for constructing confidence intervals. If the observed counts don’t match the expected, the statistic will be large. It can’t be “too small.” That would just mean that our model *really* fit the data well. So the chi-square test is always one-sided. If the calculated statistic value is large enough, we’ll reject the null hypothesis. What could be simpler?

Even though its mechanics work like a one-sided test, the interpretation of a chi-square test is in some sense *many*-sided. With more than two proportions, there are many ways the null hypothesis can be wrong. By squaring the differences, we made all the deviations positive, whether our observed counts were higher or lower than expected. There’s no direction to the rejection of the null model. All we know is that it doesn’t fit.

**FOR EXAMPLE****Doing a goodness-of-fit test**

**Recap:** We’re looking at data on the birth months of major league baseball players. We’ve checked the assumptions and conditions for performing a  $\chi^2$  test.

**Questions:** What are the hypotheses, and what does the test show?

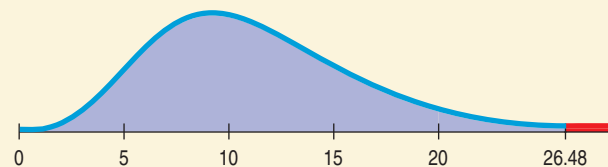
$H_0$ : The distribution of birth months for major league ballplayers is the same as that for the general population.

$H_A$ : The distribution of birth months for major league ballplayers differs from that of the rest of the population.

$$\begin{aligned} df &= 12 - 1 = 11 \\ \chi^2 &= \sum \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} \\ &= \frac{(137 - 118.24)^2}{118.24} + \frac{(121 - 103.46)^2}{103.46} + \dots \\ &= 26.48 \text{ (by technology)} \end{aligned}$$

$$P\text{-value} = P(\chi^2_{11} \geq 26.48) = 0.0055 \text{ (by technology)}$$

Because of the small P-value, I reject  $H_0$ ; there’s evidence that birth months of major league ballplayers have a different distribution from the rest of us.



## STEP-BY-STEP EXAMPLE

## A Chi-Square Test for Goodness-of-Fit

We have counts of 256 executives in 12 zodiac sign categories. The natural null hypothesis is that birth dates of executives are divided equally among all the zodiac signs. The test statistic looks at how closely the observed data match this idealized situation.

**Question:** Are zodiac signs of CEOs distributed uniformly?

THINK

**Plan** State what you want to know.

Identify the variables and check the W's.

**Hypotheses** State the null and alternative hypotheses. For  $\chi^2$  tests, it's usually easier to do that in words than in symbols.

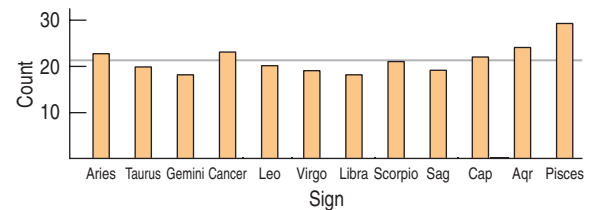
**Model** Make a picture. The null hypothesis is that the frequencies are equal, so a bar chart (with a line at the hypothesized "equal" value) is a good display.

Think about the assumptions and check the conditions.

I want to know whether births of successful people are uniformly distributed across the signs of the zodiac. I have counts of 256 Fortune 400 executives, categorized by their birth sign.

$H_0$ : Births are uniformly distributed over zodiac signs.<sup>2</sup>

$H_A$ : Births are not uniformly distributed over zodiac signs.



The bar chart shows some variation from sign to sign, and Pisces is the most frequent. But it is hard to tell whether the variation is more than I'd expect from random variation.

- ✓ **Counted Data Condition:** I have counts of the number of executives in 12 categories.
- ✓ **Independence Assumption:** The birth dates of executives should be independent of each other.
- ✓ **Randomization Condition:** This is a convenience sample of executives, but there's no reason to suspect bias.
- ✓ **Expected Cell Frequency Condition:** The null hypothesis expects that 1/12 of the 256 births, or 21.333, should occur in each sign. These expected values are all at least 5, so the condition is satisfied.

<sup>2</sup> It may seem that we have broken our rule of thumb that null hypotheses should specify parameter values. If you want to get formal about it, the null hypothesis is that

$$p_{\text{Aries}} = p_{\text{Taurus}} = \cdots = p_{\text{Pisces}}$$

That is, we hypothesize that the true proportions of births of CEOs under each sign are equal. The role of the null hypothesis is to specify the model so that we can compute the test statistic. That's what this one does.



Specify the sampling distribution model.

Name the test you will use.

The conditions are satisfied, so I'll use a  $\chi^2$  model with  $12 - 1 = 11$  degrees of freedom and do a **chi-square goodness-of-fit test**.

**SHOW**

**Mechanics** Each cell contributes an  $\frac{(Obs - Exp)^2}{Exp}$  value to the chi-square sum.

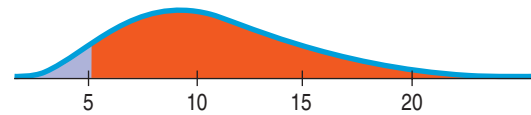
We add up these components for each zodiac sign. If you do it by hand, it can be helpful to arrange the calculation in a table. We show that after this Step-By-Step.

The P-value is the area in the upper tail of the  $\chi^2$  model above the computed  $\chi^2$  value.

The  $\chi^2$  models are skewed to the high end, and change shape depending on the degrees of freedom. The P-value considers only the right tail. Large  $\chi^2$  statistic values correspond to small P-values, which lead us to reject the null hypothesis.

The expected value for each zodiac sign is 21.333.

$$\begin{aligned}\chi^2 &= \sum \frac{(Obs - Exp)^2}{Exp} = \frac{(23 - 21.333)^2}{21.333} \\ &\quad + \frac{(20 - 21.333)^2}{21.333} + \dots \\ &= 5.094 \text{ for all 12 signs.}\end{aligned}$$



$$P\text{-value} = P(\chi^2 > 5.094) = 0.926$$

**TELL**

**Conclusion** Link the P-value to your decision. Remember to state your conclusion in terms of what the data mean, rather than just making a statement about the distribution of counts.

The P-value of 0.926 says that if the zodiac signs of executives were in fact distributed uniformly, an observed chi-square value of 5.09 or higher would occur about 93% of the time. This certainly isn't unusual, so I fail to reject the null hypothesis, and conclude that these data show virtually no evidence of nonuniform distribution of zodiac signs among executives.

## The Chi-Square Calculation

**AS**

**Activity: Calculating Standardized Residuals.** Women were at risk, too. Standardized residuals help us understand the relative risks.

Let's make the chi-square procedure very clear. Here are the steps:

1. **Find the expected values.** These come from the null hypothesis model. Every model gives a hypothesized proportion for each cell. The expected value is the product of the total number of observations times this proportion.

For our example, the null model hypothesizes *equal* proportions. With 12 signs,  $1/12$  of the 256 executives should be in each category. The expected number for each sign is 21.333.

2. **Compute the residuals.** Once you have expected values for each cell, find the residuals, *Observed* - *Expected*.
3. **Square the residuals.**
4. **Compute the components.** Now find the component,  $\frac{(Observed - Expected)^2}{Expected}$ , for each cell.

**AS** **Activity: The Chi-Square Test.** This animation completes the calculation of the chi-square statistic and the hypothesis test based on it.

5. **Find the sum of the components.** That’s the chi-square statistic.
6. **Find the degrees of freedom.** It’s equal to the number of cells minus one. For the zodiac signs, that’s  $12 - 1 = 11$  degrees of freedom.
7. **Test the hypothesis.** Large chi-square values mean lots of deviation from the hypothesized model, so they give small P-values. Look up the critical value from a table of chi-square values, or use technology to find the P-value directly.

The steps of the chi-square calculations are often laid out in tables. Use one row for each category, and columns for observed counts, expected counts, residuals, squared residuals, and the contributions to the chi-square total like this:

Sign	Observed	Expected	Residual = (Obs - Exp)	(Obs - Exp) <sup>2</sup>	Component = $\frac{(Obs - Exp)^2}{Exp}$
Aries	23	21.333	1.667	2.778889	0.130262
Taurus	20	21.333	-1.333	1.776889	0.083293
Gemini	18	21.333	-3.333	11.108889	0.520737
Cancer	23	21.333	1.667	2.778889	0.130262
Leo	20	21.333	-1.333	1.776889	0.083293
Virgo	19	21.333	-2.333	5.442889	0.255139
Libra	18	21.333	-3.333	11.108889	0.520737
Scorpio	21	21.333	-0.333	0.110889	0.005198
Sagittarius	19	21.333	-2.333	5.442889	0.255139
Capricorn	22	21.333	0.667	0.444889	0.020854
Aquarius	24	21.333	2.667	7.112889	0.333422
Pisces	29	21.333	7.667	58.782889	2.755491
					$\Sigma = 5.094$

**TI Tips**

**Testing goodness of fit**

As always, the TI makes doing the mechanics of a goodness-of-fit test pretty easy, but it does take a little work to set it up. Let’s use the zodiac data to run through the steps for a  $\chi^2$  GOF-Test.

- Enter the counts of executives born under each star sign in **L1**.  
Those counts were: 23 20 18 23 20 19 18 21 19 22 24 29
- Enter the expected percentages (or fractions, here 1/12) in **L2**. In this example they are all the same value, but that’s not always the case.
- Convert the expected percentages to expected counts by multiplying each of them by the total number of observations. We use the calculator’s summation command in the **LIST MATH** menu to find the total count for the data summarized in **L1** and then multiply that sum by the percentages stored in **L2** to produce the expected counts. The command is `sum(L1)*L2 → L2`. (We don’t ever need the percentages again, so we can replace them by storing the expected counts in **L2** instead.)
- Choose **D:  $\chi^2$  GOF-Test** from the **STATS TESTS** menu.
- Specify the lists where you stored the observed and expected counts, and enter the number of degrees of freedom, here 11.

L1	L2	L3	2
23	.083333	-----	
20	.083333		
18	.083333		
23	.083333		
20	.083333		
19	.083333		
18	.083333		

L2(5) = 1/12

SUM(L1)\*L2 → L2  
(21.33333333 21...  
(L1-L2)^2/L2 → L3  
{.1302083333 .0...

L1	L2	L3	2
23	21.333	.13021	
20	21.333	.08329	
18	21.333	.52073	
23	21.333	.13021	
20	21.333	.08329	
19	21.333	.25513	
18	21.333	.52073	

L2(7) = 21.3333333...

```

χ²GOF-Test
Observed:L1
Expected:L2
df:11
Calculate Draw
    
```

```

χ²GOF-Test
χ²=5.09375
P=.9265413914
df=11
CINTRB=C.130208...
    
```

LE	LG	CINTRB ?
-----	-----	.08333
		.08333
		.52083
		.13021
		.08333
		.25811
		.52083

CINTRB(C)=.13020833...

```

χ²cdf(5.09375,99
9,11)
.9265413914
    
```

- Ready, set, **Calculate** . . .
- . . . and there are the calculated value of  $\chi^2$  and your P-value.
- Notice, too, there's a list of values called **CINTRB**. You can scroll across them, or use **LIST NAMES** to display them as a data list (as seen on the next page). Those are the cell-by-cell components of the  $\chi^2$  calculation. We aren't very interested in them this time, because our data failed to provide evidence that the zodiac sign mattered. However, in a situation where we rejected the null hypothesis, we'd want to look at the components to see where the biggest effects occurred. You'll read more about doing that later in this chapter.

**By hand?**

If there are only a few cells, you may find that it's just as easy to write out the formula and then simply use the calculator to help you with the arithmetic. After you have found  $\chi^2 = 5.09375$  you can use your TI to find the P-value, the probability of observing a  $\chi^2$  value at least as high as the one you calculated from your data. As you probably expect, that process is akin to **normalcdf** and **tcdf**. You'll find what you need in the **DISTR** menu at **8:χ²cdf**. Just specify the left and right boundaries and the number of degrees of freedom.

- Enter  $\chi^2cdf(5.09375, 999, 11)$ , as shown. (Why 999? Unlike  $t$  and  $z$ , chi-square values can get pretty big, especially when there are many cells. You may need to go a long way to the right to get to where the curve's tail becomes essentially meaningless. You can see what we mean by looking at Table C, showing chi-square values.)

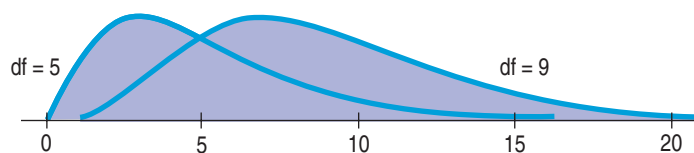
And there's the P-value, a whopping 0.93! There's nothing at all unusual about these data. (So much for the zodiac's predictive power.)

**A S** **Lesson: The Chi-Square Family of Curves.** (Not an activity like the others, but there's no better way to see how  $\chi^2$  changes with more df.) Click on the Lesson Book's Resources tab and open the chi-square table. Watch the curve at the top as you click on a row and scroll down the degrees-of-freedom column.

**How big is big?** When we calculated  $\chi^2$  for the zodiac sign example, we got 5.094. That value would have been big for  $z$  or  $t$ , leading us to reject the null hypothesis. Not here. Were you surprised that  $\chi^2 = 5.094$  had a huge P-value of 0.926? What *is* big for a  $\chi^2$  statistic, anyway?

Think about how  $\chi^2$  is calculated. In every cell, any deviation from the expected count contributes to the sum. Large deviations generally contribute more, but if there are a lot of cells, even small deviations can add up, making the  $\chi^2$  value larger. So the more cells there are, the higher the value of  $\chi^2$  has to get before it becomes noteworthy. For  $\chi^2$ , then, the decision about how big is big depends on the number of degrees of freedom.

Unlike the Normal and  $t$  families,  $\chi^2$  models are skewed. Curves in the  $\chi^2$  family change both shape and center as the number of degrees of freedom grows. Here, for example, are the  $\chi^2$  curves for 5 and 9 degrees of freedom.



Notice that the value  $\chi^2 = 10$  might seem somewhat extreme when there are 5 degrees of freedom, but appears to be rather ordinary for 9 degrees of freedom. Here are two simple facts to help you think about  $\chi^2$  models:

- ▶ The mode is at  $\chi^2 = df - 2$ . (Look back at the curves; their peaks are at 3 and 7, see?)
- ▶ The expected value (mean) of a  $\chi^2$  model is its number of degrees of freedom. That's a bit to the right of the mode—as we would expect for a skewed distribution.

Our test for zodiac birthdays had 11 df, so the relevant  $\chi^2$  curve peaks at 9 and has a mean of 11. Knowing that, we might have easily guessed that the calculated  $\chi^2$  value of 5.094 wasn't going to be significant.

## But I Believe the Model . . .



Goodness-of-fit tests are likely to be performed by people who have a theory of what the proportions *should* be in each category and who believe their theory to be true. Unfortunately, the only *null* hypothesis available for a goodness-of-fit test is that the theory is true. And as we know, the hypothesis-testing procedure allows us only to *reject* the null or *fail to reject* it. We can never confirm that a theory is in fact true, which is often what people want to do.

Unfortunately, they're stuck. At best, we can point out that the data are consistent with the proposed theory. But this doesn't *prove* the theory. The data *could* be consistent with the model even if the theory were wrong. In that case, we fail to reject the null hypothesis but can't conclude anything for sure about whether the theory is true.

And we can't fix the problem by turning things around. Suppose we try to make our favored hypothesis the alternative. Then it is impossible to pick a single null. For example, suppose, as a doubter of astrology, you want to prove that the distribution of executive births is uniform. If you choose uniform as the null hypothesis, you can only *fail* to reject it. So you'd like uniformity to be your alternative hypothesis. Which particular violation of equally distributed births would you choose as your null? The problem is that the model can be wrong in many, many ways. There's no way to frame a null hypothesis the other way around. There's just no way to prove that a favored model is true.



**Why can't we prove the null?** A biologist wanted to show that her inheritance theory about fruit flies is valid. It says that 10% of the flies should be type 1, 70% type 2, and 20% type 3. After her students collected data on 100 flies, she did a goodness-of-fit test and found a P-value of 0.07. She started celebrating, since her null hypothesis wasn't rejected—that is, until her students collected data on 100 more flies. With 200 flies, the P-value dropped to 0.02. Although she knew the answer was probably no, she asked the statistician somewhat hopefully if she could just ignore half the data and stick with the original 100. By this reasoning we could always “prove the null” just by not collecting much data. With only a little data, the chances are good that they'll be consistent with almost anything. But they also have little chance of disproving anything either. In this case, the test has no power. Don't let yourself be lured into this scientist's reasoning. With data, more is always better. But you can't ever prove that your null hypothesis is true.

## Comparing Observed Distributions

Many colleges survey graduating classes to determine the plans of the graduates. We might wonder whether the plans of students are the same at different colleges. Here's a **two-way table** for Class of 2006 graduates from several colleges at one university. Each **cell** of the table shows how many students from a particular college made a certain choice.

**WHO** Graduates from 4 colleges at an upstate New York university

**WHAT** Post-graduation activities

**WHEN** 2006

**WHY** Survey for general information

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
<b>Employed</b>	379	305	243	125	<b>1052</b>
<b>Grad School</b>	186	238	202	96	<b>722</b>
<b>Other</b>	104	123	37	58	<b>322</b>
<b>Total</b>	<b>669</b>	<b>666</b>	<b>482</b>	<b>279</b>	<b>2096</b>

**Table 26.1** Post-graduation activities of the class of 2006 for several colleges of a large university.

Because class sizes are so different, we see differences better by examining the proportions for each class rather than the counts:

**A S** **Video: The Incident.** You may have guessed which famous incident put women and children at risk. Here you can view the story complete with rare film footage.

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
<b>Employed</b>	56.7%	45.8%	50.4%	44.8%	<b>50.2</b>
<b>Grad School</b>	27.8	35.7	41.9	34.4	<b>34.4</b>
<b>Other</b>	15.5	18.5	7.7	20.8	<b>15.4</b>
<b>Total</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>

**Table 26.2** Activities of graduates as a percentage of respondents from each college.

We already know how to test whether *two* proportions are the same. For example, we could use a two-proportion *z*-test to see whether the proportion of students choosing graduate school is the same for Agriculture students as for Engineering students. But now we have more than two groups. We want to test whether the students' choices are the same across all four colleges. **The *z*-test for two proportions generalizes to a chi-square test of homogeneity.**

Chi-square again? It turns out that the mechanics of this test are *identical* to the chi-square test for goodness-of-fit that we just saw. (How similar can you get?) Why a different name, then? The goodness-of-fit test compared counts with a theoretical model. But here we're asking whether choices are the same among different groups, so we find the expected counts for each category directly from the data. As a result, we count the degrees of freedom slightly differently as well.

The term "homogeneity" means that things are the same. Here, we ask whether the post-graduation choices made by students are the *same* for these four colleges. The homogeneity test comes with a built-in null hypothesis: We hypothesize that the distribution does not change from group to group. The test looks for differences large enough to step beyond what we might expect from random sample-to-sample variation. It can reveal a large deviation in a single category or small, but persistent, differences over all the categories—or anything in between.

## Assumptions and Conditions

The assumptions and conditions are the same as for the chi-square test for goodness-of-fit. The **Counted Data Condition** says that these data must be counts. You can't do a test of homogeneity on proportions, so we have to work with the counts of graduates given in the first table. Also, you can't do a chi-square test on measurements. For example, if we had recorded GPAs for these same groups,

we wouldn't be able to determine whether the mean GPAs were different using this test.<sup>3</sup>

Often when we test for homogeneity, we aren't interested in some larger population, so we don't really need a random sample. (We would need one if we wanted to draw a more general conclusion—say, about the choices made by all members of the Class of '06.) Don't we need *some* randomness, though? Fortunately, the null hypothesis can be thought of as a model in which the counts in the table are distributed as if each student chose a plan randomly according to the overall proportions of the choices, regardless of the student's class. As long as we don't want to generalize, we don't have to check the **Randomization Condition** or the **10% Condition**.

We still must be sure we have enough data for this method to work. The **Expected Cell Frequency Condition** says that the expected count in each cell must be at least 5. We'll confirm that as we do the calculations.

## Calculations



The null hypothesis says that the proportions of graduates choosing each alternative should be the same for all four colleges, so we can estimate those overall proportions by pooling our data from the four colleges together. Within each college, the expected proportion for each choice is just the overall proportion of all students making that choice. The expected counts are those proportions applied to the number of students in each graduating class.

For example, overall, 1052, or about 50.2%, of the 2096 students who responded to the survey were employed. If the distributions are homogeneous (as the null hypothesis asserts), then 50.2% of the 669 Agriculture school graduates (or about 335.8 students) should be employed. Similarly, 50.2% of the 482 Engineering grads (or about 241.96) should be employed.

Working in this way, we (or, more likely, the computer) can fill in expected values for each cell. Because these are theoretical values, they don't have to be integers. The expected values look like this:

	Agriculture	Arts & Sciences	Engineering	Social Science	Total
Employed	335.777	334.271	241.920	140.032	1052
Grad School	230.448	229.414	166.032	96.106	722
Other	102.776	102.315	74.048	42.862	322
Total	669	666	482	279	2096

**Table 26.3** Expected values for the '06 graduates.

Now check the **Expected Cell Frequency Condition**. Indeed, there are at least 5 individuals expected in each cell.

Following the pattern of the goodness-of-fit test, we compute the component for each cell of the table. For the highlighted cell, employed students graduating from the Ag school, that's

$$\frac{(Obs - Exp)^2}{Exp} = \frac{(379 - 335.777)^2}{335.777} = 5.564$$

<sup>3</sup> To do that, you'd use a method called Analysis of Variance, discussed in a supplementary chapter on the DVD and in ActivStats.

Summing these components across all cells gives

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} = 54.51$$

How about the degrees of freedom? We don't really need to calculate all the expected values in the table. We know there is a total of 1052 employed students, so once we find the expected values for three of the colleges, we can determine the expected number for the fourth by just subtracting. Similarly, we know how many students graduated from each college, so after filling in three rows, we can find the expected values for the remaining row by subtracting. To fill out the table, we need to know the counts in only  $R - 1$  rows and  $C - 1$  columns. So the table has  $(R - 1)(C - 1)$  degrees of freedom.

In our example, we need to calculate only 2 choices in each column and counts for 3 of the 4 colleges, for a total of  $2 \times 3 = 6$  degrees of freedom. We'll need the degrees of freedom to find a P-value for the chi-square statistic.

**NOTATION ALERT:**

For a contingency table,  $R$  represents the number of rows and  $C$  the number of columns.

**STEP-BY-STEP EXAMPLE**

**A Chi-Square Test for Homogeneity**

We have reports from four colleges on the post-graduation activities of their 2006 graduating classes.

**Question:** Are students' choices of post-graduation activities the same across all the colleges?



**Plan** State what you want to know. Identify the variables and check the W's.

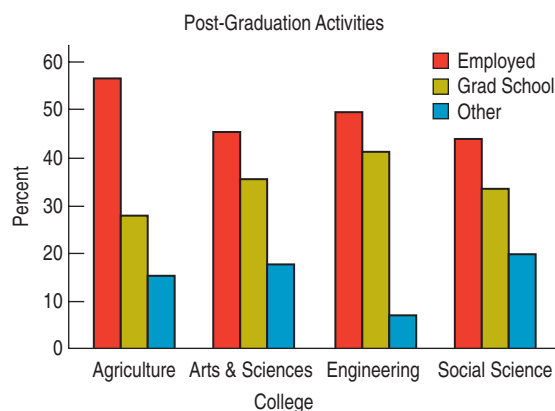
**Hypotheses** State the null and alternative hypotheses.

**Model** Make a picture: A side-by-side bar chart shows the four distributions of post-graduation activities. Plot column percents to remove the effect of class size differences. A split bar chart would also be an appropriate choice.

I want to know whether post-graduation choices are the same for students from each of four colleges. I have a table of counts classifying each college's Class of 2006 respondents according to their activities.

$H_0$ : Students' post-graduation activities are distributed in the same way for all four colleges.

$H_A$ : Students' plans do not have the same distribution.



A side-by-side bar chart shows how the distributions of choices differ across the four colleges.

Think about the assumptions and check the conditions.

State the sampling distribution model and name the test you will use.

- ✓ **Counted Data Condition:** I have counts of the number of students in categories.
- ✓ **Independence Assumption:** Student plans should be largely independent of each other. The occasional friends who decide to join Teach for America together or couples who make grad school decisions together are too rare to affect this analysis.
- ✓ **Randomization Condition:** I don't want to draw inferences to other colleges or other classes, so there is no need to check for a random sample.
- ✓ **Expected Cell Frequency Condition:** The expected values (shown below) are all at least 5.

The conditions seem to be met, so I can use a  $\chi^2$  model with  $(3 - 1) \times (4 - 1) = 6$  degrees of freedom and do a **chi-square test of homogeneity**.

**SHOW**

**Mechanics** Show the expected counts for each cell of the data table. You could make separate tables for the observed and expected counts, or put both counts in each cell as shown here. While observed counts must be whole numbers, expected counts rarely are—don't be tempted to round those off.

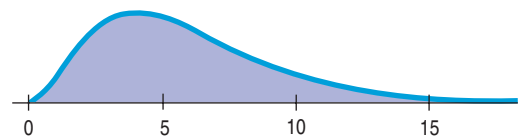
Calculate  $\chi^2$ .

The shape of a  $\chi^2$  model depends on the degrees of freedom. A  $\chi^2$  model with 6 df is skewed to the high end.

The P-value considers only the right tail. Here, the calculated value of the  $\chi^2$  statistic is off the scale, so the P-value is quite small.

	Ag	A&S	Eng	Soc Sci
Empl.	379 335.777	305 334.271	243 241.920	125 140.032
Grad sch.	186 230.448	238 229.414	202 166.032	96 96.106
Other	104 102.776	123 102.315	37 74.048	58 42.862

$$\begin{aligned} \chi^2 &= \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} \\ &= \frac{(379 - 335.777)^2}{335.777} + \dots \\ &= 54.52 \end{aligned}$$



$$P\text{-value} = P(\chi^2 > 54.52) < 0.0001$$

**TELL**

**Conclusion** State your conclusion in the context of the data. You should specifically talk about whether the distributions for the groups appear to be different.

The P-value is very small, so I reject the null hypothesis and conclude that there's evidence that the post-graduation activities of students from these four colleges don't have the same distribution.



If you find that simply rejecting the hypothesis of homogeneity is a bit unsatisfying, you're in good company. Ok, so the post-graduation plans are different. What we'd really like to know is what the differences are, where they're the greatest, and where they're smallest. The test for homogeneity doesn't answer these interesting questions, but it does provide some evidence that can help us.

## Examining the Residuals

Whenever we reject the null hypothesis, it's a good idea to examine residuals. (We don't need to do that when we fail to reject because when the  $\chi^2$  value is small, all of its components must have been small.) For chi-square tests, we want to compare residuals for cells that may have very different counts. So we're better off standardizing the residuals. We know the mean residual is zero,<sup>4</sup> but we need to know each residual's standard deviation. When we tested proportions, we saw a link between the expected proportion and its standard deviation. For counts, there's a similar link. To standardize a cell's residual, we just divide by the square root of its expected value:

$$c = \frac{(Obs - Exp)}{\sqrt{Exp}}$$

Notice that these **standardized residuals** are just the square roots of the **components** we calculated for each cell, and their sign indicates whether we observed more cases than we expected, or fewer.

The standardized residuals give us a chance to think about the underlying patterns and to consider the ways in which the distribution of post-graduation plans may differ from college to college. Now that we've subtracted the mean (zero) and divided by their standard deviations, these are z-scores. If the null hypothesis were true, we could even appeal to the Central Limit Theorem, think of the Normal model, and use the 68–95–99.7 Rule to judge how extraordinary the large ones are.

Here are the standardized residuals for the Class of '06 data:

	Ag	A&S	Eng	Soc Sci
Employed	2.359	-1.601	0.069	-1.270
Grad School	-2.928	0.567	2.791	-0.011
Other	0.121	2.045	-4.305	2.312

**Table 26.4**

Standardized residuals can help show how the table differs from the null hypothesis pattern.

The column for Engineering students immediately attracts our attention. It holds both the largest positive and the largest negative standardized residuals. It looks like Engineering college graduates are more likely to go on to graduate work and very unlikely to take time off for "volunteering and travel, among other activities" (as the "Other" category is explained). By contrast, Ag school graduates seem to be readily employed and less likely to pursue graduate work immediately after college.

<sup>4</sup> Residual = observed – expected. Because the total of the expected values is set to be the same as the observed total, the residuals must sum to zero.

## FOR EXAMPLE

Looking at  $\chi^2$  residuals

**Recap:** Some people suggest that school children who are the older ones in their class naturally perform better in sports and therefore get more coaching and encouragement. To see if there's any evidence for this, we looked at major league baseball players born since 1975. A goodness-of-fit test found their birth months to have a distribution that's significantly different from the rest of us. The table shows the standardized residuals.

**Question:** What's different about the distribution of birth months among major league ballplayers?

It appears that, compared to the general population, fewer ballplayers than expected were born in July and more than expected in August. Either month would make them the younger kids in their grades in school, so these data don't offer support for the conjecture that being older is an advantage in terms of a career as a pro athlete.

Month	Residual	Month	Residual
1	1.73	7	-2.69
2	1.72	8	2.77
3	-0.21	9	0.08
4	0.25	10	-1.56
5	0.71	11	-1.22
6	-0.39	12	-0.96



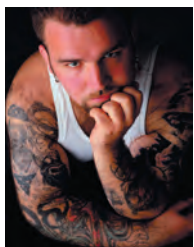
## JUST CHECKING

Tiny black potato flea beetles can damage potato plants in a vegetable garden. These pests chew holes in the leaves, causing the plants to wither or die. They can be killed with an insecticide, but a canola oil spray has been suggested as a non-chemical “natural” method of controlling the beetles. To conduct an experiment to test the effectiveness of the natural spray, we gather 500 beetles and place them in three Plexiglas® containers. Two hundred beetles go in the first container, where we spray them with the canola oil mixture. Another 200 beetles go in the second container; we spray them with the insecticide. The remaining 100 beetles in the last container serve as a control group; we simply spray them with water. Then we wait 6 hours and count the number of surviving beetles in each container.

1. Why do we need the control group?
2. What would our null hypothesis be?
3. After the experiment is over, we could summarize the results in a table as shown. How many degrees of freedom does our  $\chi^2$  test have?
4. Suppose that, all together, 125 beetles survived. (That's the first-row total.) What's the expected count in the first cell—survivors among those sprayed with the natural spray?
5. If it turns out that only 40 of the beetles in the first container survived, what's the calculated component of  $\chi^2$  for that cell?
6. If the total calculated value of  $\chi^2$  for this table turns out to be around 10, would you expect the P-value of our test to be large or small? Explain.

	Natural spray	Insecticide	Water	Total
Survived				
Died				
Total	200	200	100	500

## Independence



A study from the University of Texas Southwestern Medical Center examined whether the risk of hepatitis C was related to whether people had tattoos and to where they got their tattoos. Hepatitis C causes about 10,000 deaths each year in the United States, but often lies undetected for years after infection.

The data from this study can be summarized in a two-way table, as follows:

**WHO** Patients being treated for non-blood-related disorders

**WHAT** Tattoo status and hepatitis C status

**WHEN** 1991, 1992

**WHERE** Texas

	Hepatitis C	No Hepatitis C	Total
Tattoo, parlor	17	35	52
Tattoo, elsewhere	8	53	61
None	22	491	513
Total	47	579	626

**Table 26.5**

Counts of patients classified by their hepatitis C test status according to whether they had a tattoo from a tattoo parlor or from another source, or had no tattoo.

These data differ from the kinds of data we've considered before in this chapter because they categorize subjects from a single group on two categorical variables rather than on only one. The categorical variables here are *Hepatitis C Status* ("Hepatitis C" or "No Hepatitis C") and *Tattoo Status* ("Parlor," "Elsewhere," "None"). We've seen counts classified by two categorical variables displayed like this in Chapter 3, so we know such tables are called contingency tables. **Contingency tables** categorize counts on two (or more) variables so that we can see whether the distribution of counts on one variable is contingent on the other.

The natural question to ask of these data is whether the chance of having hepatitis C is *independent* of tattoo status. Recall that for events **A** and **B** to be independent  $P(\mathbf{A})$  must equal  $P(\mathbf{A} | \mathbf{B})$ . Here, this means the probability that a randomly selected patient has hepatitis C should not change when we learn the patient's tattoo status. We examined the question of independence in just this way back in Chapter 15, but we lacked a way to test it. The rules for independent events are much too precise and absolute to work well with real data. **A chi-square test for independence** is called for here.

If *Hepatitis Status* is independent of tattoos, we'd expect the proportion of people testing positive for hepatitis to be the same for the three levels of *Tattoo Status*. This sounds a lot like the test of homogeneity. In fact, the mechanics of the calculation are identical.

The difference is that now we have two categorical variables measured on a single population. For the homogeneity test, we had a single categorical variable measured independently on two or more populations. But now we ask a different question: "Are the variables independent?" rather than "Are the groups homogeneous?" These are subtle differences, but they are important when we state hypotheses and draw conclusions.

**A S** **Activity: Independence and Chi-Square.** This unusual simulation shows how independence arises (and fails) in contingency tables.

The only difference between the test for homogeneity and the test for independence is in what you . . .

**THINK**

**FOR EXAMPLE** Which  $\chi^2$  test?

Many states and localities now collect data on traffic stops regarding the race of the driver. The initial concern was that Black drivers were being stopped more often (the "crime" ironically called "Driving While Black"). With more data in hand, attention has turned to other issues. For example, data from 2533 traffic stops in Cincinnati<sup>5</sup> report the race of the driver (Black, White, or Other) and whether the traffic stop resulted in a search of the vehicle.

		Race			Total
		Black	White	Other	
Search	No	787	594	27	1408
	Yes	813	293	19	1125
	Total	1600	887	46	2533

**Question:** Which test would be appropriate to examine whether race is a factor in vehicle searches? What are the hypotheses?

(continued)

<sup>5</sup> John E. Eck, Lin Liu, and Lisa Growette Bostaph, *Police Vehicle Stops in Cincinnati*, Oct. 1, 2003, available at <http://www.cincinnati-oh.gov>. Data for other localities can be found by searching from <http://www.racialprofilinganalysis.neu.edu>.

For Example (*continued*)

These data represent one group of traffic stops in Cincinnati, categorized on two variables, Race and Search. I'll do a chi-square test of independence.

$H_0$ : Whether or not police search a vehicle is independent of the race of the driver.

$H_A$ : Decisions to search vehicles are not independent of the driver's race.

## Assumptions and Conditions

**A S** **Activity: Chi-Square Tables.** Work with *ActivStats*' interactive chi-square table to perform a hypothesis test.

Of course, we still need counts and enough data so that the expected values are at least 5 in each cell.

If we're interested in the independence of variables, we usually want to generalize from the data to some population. In that case, we'll need to check that the data are a representative random sample from, and fewer than 10% of, that population.

### STEP-BY-STEP EXAMPLE

### A Chi-Square Test for Independence

We have counts of 626 individuals categorized according to their "tattoo status" and their "hepatitis status."

**Question:** Are tattoo status and hepatitis status independent?

**THINK**

**Plan** State what you want to know.

Identify the variables and check the W's.

**Hypotheses** State the null and alternative hypotheses.

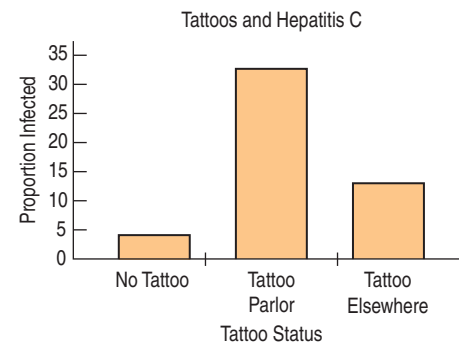
We perform a test of independence when we suspect the variables may not be independent. We are on the familiar ground of making a claim (in this case, that knowing *Tattoo Status* will change probabilities for *Hepatitis C Status*) and testing the null hypothesis that it is *not* true.

**Model** Make a picture. Because these are only two categories—Hepatitis C and No Hepatitis C—a simple bar chart of the distribution of tattoo sources for Hep C patients shows all the information.

I want to know whether the categorical variables *Tattoo Status* and *Hepatitis Status* are statistically independent. I have a contingency table of 626 Texas patients with an unrelated disease.

$H_0$ : *Tattoo Status* and *Hepatitis Status* are independent.<sup>6</sup>

$H_A$ : *Tattoo Status* and *Hepatitis Status* are not independent.



The bar chart suggests strong differences in Hepatitis C risk based on tattoo status.

<sup>6</sup> Once again, parameters are hard to express. The hypothesis of independence itself tells us how to find expected values for each cell of the contingency table. That's all we need.

Think about the assumptions and check the conditions.

This table shows both the observed and expected counts for each cell. The expected counts are calculated exactly as they were for a test of homogeneity; in the first cell, for example, we expect  $\frac{52}{626}$  (that's 8.3%) of 47.

*Warning:* Be wary of proceeding when there are small expected counts, If we see expected counts that fall far short of 5, or if many cells violate the condition, we should not use  $\chi^2$ . (We will soon discuss ways you can fix the problem.) If you do continue, always check the residuals to be sure those cells did not have a major influence on your result.

Specify the model.

Name the test you will use.

- ✓ **Counted Data Condition:** I have counts of individuals categorized on two variables.
- ✓ **Independence Assumption:** The people in this study are likely to be independent of each other.
- ✓ **Randomization Condition:** These data are from a retrospective study of patients being treated for something unrelated to hepatitis. Although they are not an SRS, they were selected to avoid biases.
- ✓ **10% Condition:** These 626 patients are far fewer than 10% of all those with tattoos or hepatitis.
- ✗ **Expected Cell Frequency Condition:** The expected values do not meet the condition that all are at least 5.

	Hepatitis C	No Hepatitis C	Total
Tattoo, parlor	17 3.904	35 48.096	52
Tattoo, elsewhere	8 4.580	53 56.420	61
None	22 38.516	491 474.484	513
Total	47	579	626

Although the Expected Cell Frequency Condition is not satisfied, the values are close to 5. I'll go ahead, but I'll check the residuals carefully. I'll use a  $\chi^2$  model with  $(3 - 1) \times (2 - 1) = 2$  df and do a **chi-square test of independence**.

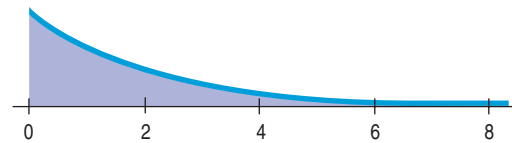


**Mechanics** Calculate  $\chi^2$ .

The shape of a chi-square model depends on its degrees of freedom. With 2 df, the model looks quite different, as you can

$$\begin{aligned} \chi^2 &= \sum_{\text{all cells}} \frac{(Obs - Exp)^2}{Exp} \\ &= \frac{(17 - 3.904)^2}{3.904} + \dots = 57.91 \end{aligned}$$

see here. We still care only about the right tail.



**Conclusion** Link the P-value to your decision. State your conclusion about the independence of the two variables.

(We should be wary of this conclusion because of the small expected counts. A complete solution must include the additional analysis, recalculation, and final conclusion discussed in the following section.)

The P-value is very small, so I reject the null hypothesis and conclude that *Hepatitis Status* is not independent of *Tattoo Status*. Because the Expected Cell Frequency Condition was violated, I need to check that the two cells with small expected counts did not influence this result too greatly.

**FOR EXAMPLE**

**Chi-square mechanics**

**Recap:** We have data that allow us to investigate whether police searches of vehicles they stop are independent of the driver's race.

**Questions:** What are the degrees of freedom for this test? What is the expected frequency of searches for the Black drivers who were stopped? What's that cell's component in the  $\chi^2$  computation? And how is the standardized residual for that cell computed?

This is a  $2 \times 3$  contingency table, so  $df = (2 - 1)(3 - 1) = 2$ .

Overall, 1125 of 2533 vehicles were searched. If searches are conducted independent of race, then I'd expect  $\frac{1125}{2533}$  of the 1600 Black drivers to have been searched:  $\frac{1125}{2533} \times 1600 \approx 710.62$ .

That cell's term in the  $\chi^2$  calculation is  $\frac{(Obs - Exp)^2}{Exp} = \frac{(813 - 710.62)^2}{710.62} = 14.75$

The standardized residual for that cell is  $\frac{Obs - Exp}{\sqrt{Exp}} = \frac{813 - 710.62}{\sqrt{710.62}} = 3.84$

		Race			Total
		Black	White	Other	
Search	No	787	594	27	1408
	Yes	813	293	19	1125
	Total	1600	887	46	2533

## Examine the Residuals

Each cell of the contingency table contributes a term to the chi-square sum. As we did earlier, we should examine the residuals because we have rejected the null hypothesis. In this instance, we have an additional concern that the cells with small expected frequencies not be the ones that make the chi-square statistic large.

Our interest in the data arises from the potential for improving public health. If patients with tattoos are more likely to test positive for hepatitis C, perhaps physicians should be advised to suggest blood tests for such patients.

The standardized residuals look like this:



	Hepatitis C	No Hepatitis C
Tattoo, parlor	6.628	-1.888
Tattoo, elsewhere	1.598	-0.455
None	-2.661	0.758

**Table 26.6**

Standardized residuals for the hepatitis and tattoos data. Are any of them particularly large in magnitude?

The chi-square value of 57.91 is the sum of the squares of these six values. The cell for people with tattoos obtained in a tattoo parlor who have hepatitis C is large and positive, indicating there are more people in that cell than the null hypothesis of independence would predict. Maybe tattoo parlors are a source of infection or maybe those who go to tattoo parlors also engage in risky behavior.

The second-largest component is a negative value for those with no tattoos who test positive for hepatitis C. A negative value says that there are fewer people in this cell than independence would expect. That is, those who have no tattoos are less likely to be infected with hepatitis C than we might expect if the two variables were independent.

What about the cells with small expected counts? The formula for the chi-square standardized residuals divides each residual by the square root of the expected frequency. Too small an expected frequency can arbitrarily inflate the residual and lead to an inflated chi-square statistic. Any expected count close to the arbitrary minimum of 5 calls for checking that cell's standardized residual to be sure it is not particularly large. In this case, the standardized residual for the "Hepatitis C and Tattoo, elsewhere" cell is not particularly large, but the standardized residual for the "Hepatitis C and Tattoo, parlor" cell is large.

We might choose not to report the results because of concern with the small expected frequency. Alternatively, we could include a warning along with our report of the results. Yet another approach is to combine categories to get a larger sample size and correspondingly larger expected frequencies, if there are some categories that can be appropriately combined. Here, we might naturally combine the two rows for tattoos, obtaining a  $2 \times 2$  table:



	Hepatitis C	No Hepatitis C	Total
Tattoo	25	88	113
None	22	491	513
Total	47	579	626

**Table 26.7**

Combining the two tattoo categories gives a table with all expected counts greater than 5.

This table has expected values of at least 5 in every cell, and a chi-square value of 42.42 on 1 degree of freedom. The corresponding P-value is  $<0.0001$ .

We conclude that *Tattoo Status* and *Hepatitis C Status* are not independent. The data suggest that tattoo parlors may be a particular problem, but we haven't enough data to draw that conclusion.



## FOR EXAMPLE

Writing conclusions for  $\chi^2$  tests

**Recap:** We're looking at Cincinnati traffic stop data to see if police decisions about searching cars show evidence of racial bias. With 2 df, technology calculates  $\chi^2 = 73.25$ , a P-value less than 0.0001, and these standardized residuals:

**Question:** What's your conclusion?

The very low P-value leads me to reject the null hypothesis.

There's strong evidence that police decisions to search cars at traffic stops are associated with the driver's race.

The largest residuals are for White drivers, who are searched less often than independence would predict. It appears that Black drivers' cars are searched more often.

		Race		
		Black	White	Other
Search	No	-3.43	4.55	0.28
	Yes	3.84	-5.09	-0.31

## TI Tips

## Testing homogeneity or independence

Yes, the TI will do chi-square tests of homogeneity and independence. Let's use the tattoo data. Here goes.

**Test a hypothesis of homogeneity or independence**

Stage 1: You need to enter the data as a matrix. A "matrix" is just a formal mathematical term for a table of numbers.

- Push the **MATRIX** button, and choose to **EDIT** matrix **[A]**.
- First specify the dimensions of the table, rows  $\times$  columns.
- Enter the appropriate counts, one cell at a time. The calculator automatically asks for them row by row.

Stage 2: Do the test.

- In the **STAT TESTS** menu choose **C:  $\chi^2$ -Test**.
- The TI now confirms that you have placed the observed frequencies in **[A]**. It also tells you that when it finds the expected frequencies it will store those in **[B]** for you. Now **Calculate** the mechanics of the test.

The TI reports a calculated value of  $\chi^2 = 57.91$  and an exceptionally small P-value.

Stage 3: Check the expected counts.

- Go back to **MATRIX EDIT** and choose **[B]**.

Notice that two of the cells fail to meet the condition that expected counts be at least 5. This problem enters into our analysis and conclusions.

Stage 4: And now some bad news. There's no easy way to calculate the standardized residuals. Look at the two matrices, **[A]** and **[B]**. Large residuals will happen when the corresponding entries differ greatly, especially when the expected count in **[B]** is small (because you will divide by the square root of the entry in **[B]**). The first cell is a good candidate, so we show you the calculation of its standardized residual.

A residual of over 6 is pretty large—possibly an indication that you're more likely to get hepatitis in a tattoo parlor, but the expected count is smaller than 5. We're pretty sure that hepatitis status is not independent of having a tattoo, but we should be wary of saying anything more. Probably the best approach is to combine categories to get cells with expected counts above 5.

```
NAMES MATH 0001
[A] 2x4
[B] 2x4
[C] 3x2
[D]
[E]
[F]
[G]
```

```
MATRIX[A] 3 x2
[ 17 35 ]
[ 8 23 ]
[ 22 45 ]
3 x 2=491
```

```
EDIT CALC TESTS
B1:2-PropZInt...
C:  $\chi^2$ -Test...
D:  $\chi^2$ GOF-Test...
E: 2-SampTTest...
F: LinRegTTest...
G: LinRegInt...
H: ANOVA<
```

```
 $\chi^2$ -Test
 $\chi^2=57.91217384$ 
P=2.657855E-13
df=2
```

```
MATRIX[B] 3x2
[ 3.9042 48.096 ]
[ 4.5799 58.42 ]
[ 38.516 474.48 ]
```

```
(17-3.9042)/√(3.9042)
6.627748275
```



## Chi-Square and Causation



Chi-square tests are common. Tests for independence are especially widespread. Unfortunately, many people interpret a small P-value as proof of causation. We know better. Just as correlation between quantitative variables does not demonstrate causation, a failure of independence between two categorical variables does not show a cause-and-effect relationship between them, nor should we say that one variable *depends* on the other.

The chi-square test for independence treats the two variables symmetrically. There is no way to differentiate the direction of any possible causation from one variable to the other. In our example, it is unlikely that having hepatitis causes one to crave a tattoo, but other examples are not so clear.

In this case it's easy to imagine that lurking variables are responsible for the observed lack of independence. Perhaps the lifestyles of some people include both tattoos and behaviors that put them at increased risk of hepatitis C, such as body piercings or even drug use. Even a small subpopulation of people with such a lifestyle among those with tattoos might be enough to create the observed result. After all, we observed only 25 patients with both tattoos and hepatitis.

In some sense, a failure of independence between two categorical variables is less impressive than a strong, consistent, linear association between quantitative variables. Two categorical variables can fail the test of independence in many ways, including ways that show no consistent pattern of failure. Examination of the chi-square standardized residuals can help you think about the underlying patterns.



### JUST CHECKING

Which of the three chi-square tests—goodness-of-fit, homogeneity, or independence—would you use in each of the following situations?

7. A restaurant manager wonders whether customers who dine on Friday nights have the same preferences among the four “chef’s special” entrées as those who dine on Saturday nights. One weekend he has the wait staff record which entrées were ordered each night. Assuming these customers to be typical of all weekend diners, he’ll compare the distributions of meals chosen Friday and Saturday.
8. Company policy calls for parking spaces to be assigned to everyone at random, but you suspect that may not be so. There are three lots of equal size: lot A, next to the building; lot B, a bit farther away; and lot C, on the other side of the highway. You gather data about employees at middle management level and above to see how many were assigned parking in each lot.
9. Is a student’s social life affected by where the student lives? A campus survey asked a random sample of students whether they lived in a dormitory, in off-campus housing, or at home, and whether they had been out on a date 0, 1–2, 3–4, or 5 or more times in the past two weeks.

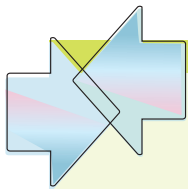
### WHAT CAN GO WRONG?

- ▶ **Don’t use chi-square methods unless you have counts.** All three of the chi-square tests apply only to counts. Other kinds of data can be arrayed in two-way tables. Just because numbers are in a two-way table doesn’t make them suitable for chi-square analysis. Data reported as proportions or percentages can be suitable for chi-square procedures, *but only after they are converted to counts*. If you try to do the calculations without first finding the counts, your results will be wrong.

(continued)

**AS** **Simulation: Sample Size and Chi-Square.** Chi-square statistics have a peculiar problem. They don't respond to increasing the sample size in quite the same way you might expect.

- ▶ **Beware large samples.** Beware *large* samples?! That's not the advice you're used to hearing. The chi-square tests, however, are unusual. You should be wary of chi-square tests performed on very large samples. No hypothesized distribution fits perfectly, no two groups are exactly homogeneous, and two variables are rarely perfectly independent. The degrees of freedom for chi-square tests don't grow with the sample size. With a sufficiently large sample size, a chi-square test can always reject the null hypothesis. But we have no measure of how far the data are from the null model. There are no confidence intervals to help us judge the effect size.
- ▶ **Don't say that one variable "depends" on the other just because they're not independent.** Dependence suggests a pattern and implies causation, but variables can fail to be independent in many different ways. When variables fail the test for independence, you might just say they are "associated."



## CONNECTIONS

Chi-square methods relate naturally to inference methods for proportions. We can think of a test of homogeneity as stepping from a comparison of two proportions to a question of whether three or more proportions are equal. The standard deviations of the residuals in each cell are linked to the expected counts much like the standard deviations we found for proportions.

Independence is, of course, a fundamental concept in Statistics. But chi-square tests do not offer a general way to check on independence for all those times when we have had to assume it.

Stacked bar charts or side-by-side pie charts can help us think about patterns in two-way tables. A histogram or boxplot of the standardized residuals can help locate extraordinary values.

## WHAT HAVE WE LEARNED?

We've learned how to test hypotheses about categorical variables. We use one of three related methods. All look at counts of data in categories, and all rely on chi-square models, a new family indexed by degrees of freedom.

- ▶ Goodness-of-fit tests compare the observed distribution of a single categorical variable to an expected distribution based on a theory or model.
- ▶ Tests of homogeneity compare the distribution of several groups for the same categorical variable.
- ▶ Tests of independence examine counts from a single group for evidence of an association between two categorical variables.

We've seen that, mechanically, these tests are almost identical. Although the tests appear to be one-sided, we've learned that conceptually they are many-sided, because there are many ways that a table of counts can deviate significantly from what we hypothesized. When that happens and we reject the null hypothesis, we've learned to examine standardized residuals in order to better understand patterns as in the table.

## Terms

**Chi-square model**

621, 625. Chi-square models are skewed to the right. They are parameterized by their degrees of freedom and become less skewed with increasing degrees of freedom.

**Cell**

619, 626. A cell is one element of a table corresponding to a specific row and a specific column. Table cells can hold counts, percentages, or measurements on other variables. Or they can hold several values.



**Chi-square statistic** 621. The chi-square statistic can be used to test whether the observed counts in a frequency distribution or contingency table match the counts we would expect according to some model. It is calculated as

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

Chi-square statistics differ in how expected counts are found, depending on the question asked.

**Chi-square test of goodness-of-fit** 618, 622. A test of whether the distribution of counts in one categorical variable matches the distribution predicted by a model is called a test of goodness-of-fit. In a chi-square goodness-of-fit test, the expected counts come from the predicting model. The test finds a P-value from a chi-square model with  $n - 1$  degrees of freedom, where  $n$  is the number of categories in the categorical variable.

**Chi-square test of homogeneity** 627. A test comparing the distribution of counts for two or more groups on the same categorical variable is called a test of *homogeneity*. A chi-square test of homogeneity finds expected counts based on the overall frequencies, adjusted for the totals in each group under the (null hypothesis) assumption that the distributions are the same for each group. We find a P-value from a chi-square distribution with  $(\#Rows - 1) \times (\#Cols - 1)$  degrees of freedom, where  $\#Rows$  gives the number of categories and  $\#Cols$  gives the number of independent groups.

**Chi-square test of independence** 633. A test of whether two categorical variables are independent examines the distribution of counts for one group of individuals classified according to both variables. A chi-square test of *independence* finds expected counts by assuming that knowing the marginal totals tells us the cell frequencies, assuming that there is no association between the variables. This turns out to be the same calculation as a test of homogeneity. We find a P-value from a chi-square distribution with  $(\#Rows - 1) \times (\#Cols - 1)$  degrees of freedom, where  $\#Rows$  gives the number of categories in one variable and  $\#Cols$  gives the number of categories in the other.

**Chi-square component** 623, 628. The components of a chi-square calculation are

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

found for each cell of the table.

**Standardized residual** 631. In each cell of a two-way table, a standardized residual is the square root of the chi-square component for that cell with the sign of the *Observed* – *Expected* difference:

$$\frac{(\text{Obs} - \text{Exp})}{\sqrt{\text{Exp}}}$$

When we reject a chi-square test, an examination of the standardized residuals can sometimes reveal more about how the data deviate from the null model.

**Two-way table** 626, 633. Each *cell* of a two-way table shows counts of individuals. One way classifies a sample according to a categorical variable. The other way can classify different groups of individuals according to the same variable or classify the same individuals according to a different categorical variable.

**Contingency table** 633. A two-way table that classifies individuals according to two categorical variables is called a *contingency table*.

## Skills



- ▶ Be able to recognize when a test of goodness-of-fit, a test of homogeneity, or a test of independence would be appropriate for a table of counts.
- ▶ Understand that the degrees of freedom for a chi-square test depend on the dimensions of the table and not on the sample size. Understand that this means that increasing the sample size increases the ability of chi-square procedures to reject the null hypothesis.



- ▶ Be able to display and interpret counts in a two-way table.
- ▶ Know how to use the chi-square tables to perform chi-square tests.

- ▶ Know how to compute a chi-square test using your statistics software or calculator.
- ▶ Be able to examine the standardized residuals to explain the nature of the deviations from the null hypothesis.



- ▶ Know how to interpret chi-square as a test of goodness-of-fit in a few sentences.
- ▶ Know how to interpret chi-square as a test of homogeneity in a few sentences.
- ▶ Know how to interpret chi-square as a test of independence in a few sentences.

## CHI-SQUARE ON THE COMPUTER

Most statistics packages associate chi-square tests with contingency tables. Often chi-square is available as an option only when you make a contingency table. This organization can make it hard to locate the chi-square test and may confuse the three different roles that the chi-square test can take. In particular, chi-square tests for goodness-of-fit may be hard to find or missing entirely. Chi-square tests for homogeneity are computationally the same as chi-square tests for independence, so you may have to perform the mechanics as if they were tests of independence and interpret them afterwards as tests of homogeneity.

Most statistics packages work with data on individuals rather than with the summary counts. If the only information you have is the table of counts, you may find it more difficult to get a statistics package to compute chi-square. Some packages offer a way to reconstruct the data from the summary counts so that they can then be passed back through the chi-square calculation, finding the cell counts again. Many packages offer chi-square standardized residuals (although they may be called something else).

## EXERCISES

1. **Which test?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
  - a) A brokerage firm wants to see whether the type of account a customer has (Silver, Gold, or Platinum) affects the type of trades that customer makes (in person, by phone, or on the Internet). It collects a random sample of trades made for its customers over the past year and performs a test.
  - b) That brokerage firm also wants to know if the type of account affects the size of the account (in dollars). It performs a test to see if the mean size of the account is the same for the three account types.
  - c) The academic research office at a large community college wants to see whether the distribution of courses chosen (Humanities, Social Science, or Science) is different for its residential and nonresidential students. It assembles last semester's data and performs a test.
2. **Which test again?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
  - a) Is the quality of a car affected by what day it was built? A car manufacturer examines a random sample of the warranty claims filed over the past two years to test whether defects are randomly distributed across days of the work week.
  - b) A medical researcher wants to know if blood cholesterol level is related to heart disease. She examines a database of 10,000 patients, testing whether the cholesterol level (in milligrams) is related to whether or not a person has heart disease.
  - c) A student wants to find out whether political leaning (liberal, moderate, or conservative) is related to choice of major. He surveys 500 randomly chosen students and performs a test.
3. **Dice.** After getting trounced by your little brother in a children's game, you suspect the die he gave you to roll may be unfair. To check, you roll it 60 times, recording the number of times each face appears. Do these results cast doubt on the die's fairness?

- ▶ Know how to compute a chi-square test using your statistics software or calculator.
- ▶ Be able to examine the standardized residuals to explain the nature of the deviations from the null hypothesis.



- ▶ Know how to interpret chi-square as a test of goodness-of-fit in a few sentences.
- ▶ Know how to interpret chi-square as a test of homogeneity in a few sentences.
- ▶ Know how to interpret chi-square as a test of independence in a few sentences.

## CHI-SQUARE ON THE COMPUTER

Most statistics packages associate chi-square tests with contingency tables. Often chi-square is available as an option only when you make a contingency table. This organization can make it hard to locate the chi-square test and may confuse the three different roles that the chi-square test can take. In particular, chi-square tests for goodness-of-fit may be hard to find or missing entirely. Chi-square tests for homogeneity are computationally the same as chi-square tests for independence, so you may have to perform the mechanics as if they were tests of independence and interpret them afterwards as tests of homogeneity.

Most statistics packages work with data on individuals rather than with the summary counts. If the only information you have is the table of counts, you may find it more difficult to get a statistics package to compute chi-square. Some packages offer a way to reconstruct the data from the summary counts so that they can then be passed back through the chi-square calculation, finding the cell counts again. Many packages offer chi-square standardized residuals (although they may be called something else).

## EXERCISES

1. **Which test?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
  - a) A brokerage firm wants to see whether the type of account a customer has (Silver, Gold, or Platinum) affects the type of trades that customer makes (in person, by phone, or on the Internet). It collects a random sample of trades made for its customers over the past year and performs a test.
  - b) That brokerage firm also wants to know if the type of account affects the size of the account (in dollars). It performs a test to see if the mean size of the account is the same for the three account types.
  - c) The academic research office at a large community college wants to see whether the distribution of courses chosen (Humanities, Social Science, or Science) is different for its residential and nonresidential students. It assembles last semester's data and performs a test.
2. **Which test again?** For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:
  - a) Is the quality of a car affected by what day it was built? A car manufacturer examines a random sample of the warranty claims filed over the past two years to test whether defects are randomly distributed across days of the work week.
  - b) A medical researcher wants to know if blood cholesterol level is related to heart disease. She examines a database of 10,000 patients, testing whether the cholesterol level (in milligrams) is related to whether or not a person has heart disease.
  - c) A student wants to find out whether political leaning (liberal, moderate, or conservative) is related to choice of major. He surveys 500 randomly chosen students and performs a test.
3. **Dice.** After getting trounced by your little brother in a children's game, you suspect the die he gave you to roll may be unfair. To check, you roll it 60 times, recording the number of times each face appears. Do these results cast doubt on the die's fairness?

- If the die is fair, how many times would you expect each face to show?
- To see if these results are unusual, will you test goodness-of-fit, homogeneity, or independence?
- State your hypotheses.
- Check the conditions.
- How many degrees of freedom are there?
- Find  $\chi^2$  and the P-value.
- State your conclusion.

Face	Count
1	11
2	7
3	9
4	15
5	12
6	6

- M&M's.** As noted in an earlier chapter, the Masterfoods Company says that until very recently yellow candies made up 20% of its milk chocolate M&M's, red another 20%, and orange, blue, and green 10% each. The rest are brown. On his way home from work the day he was writing these exercises, one of the authors bought a bag of plain M&M's. He got 29 yellow ones, 23 red, 12 orange, 14 blue, 8 green, and 20 brown. Is this sample consistent with the company's stated proportions? Test an appropriate hypothesis and state your conclusion.

  - If the M&M's are packaged in the stated proportions, how many of each color should the author have expected to get in his bag?
  - To see if his bag was unusual, should he test goodness-of-fit, homogeneity, or independence?
  - State the hypotheses.
  - Check the conditions.
  - How many degrees of freedom are there?
  - Find  $\chi^2$  and the P-value.
  - State a conclusion.
- Nuts.** A company says its premium mixture of nuts contains 10% Brazil nuts, 20% cashews, 20% almonds, and 10% hazelnuts, and the rest are peanuts. You buy a large can and separate the various kinds of nuts. Upon weighing them, you find there are 112 grams of Brazil nuts, 183 grams of cashews, 207 grams of almonds, 71 grams of hazelnuts, and 446 grams of peanuts. You wonder whether your mix is significantly different from what the company advertises.

  - Explain why the chi-square goodness-of-fit test is not an appropriate way to find out.
  - What might you do instead of weighing the nuts in order to use a  $\chi^2$  test?
- Mileage.** A salesman who is on the road visiting clients thinks that, on average, he drives the same distance each day of the week. He keeps track of his mileage for several weeks and discovers that he averages 122 miles on Mondays, 203 miles on Tuesdays, 176 miles on Wednesdays, 181 miles on Thursdays, and 108 miles on Fridays. He wonders if this evidence contradicts his belief in a uniform distribution of miles across the days of the week. Explain why it is not appropriate to test his hypothesis using the chi-square goodness-of-fit test.
- NYPD and race.** Census data for New York City indicate that 29.2% of the under-18 population is white, 28.2% black, 31.5% Latino, 9.1% Asian, and 2% other

ethnicities. The New York Civil Liberties Union points out that, of 26,181 police officers, 64.8% are white, 14.5% black, 19.1% Hispanic, and 1.4% Asian. Do the police officers reflect the ethnic composition of the city's youth? Test an appropriate hypothesis and state your conclusion.

- Violence against women 2005.** In its study *When Men Murder Women*, the Violence Policy Center ([www.vpc.org](http://www.vpc.org)) reported that 1857 women were murdered by men in 2005. Of these victims, a weapon could be identified for 1752 of them. Of those for whom a weapon could be identified, 966 were killed by guns, 390 by knives or other cutting instruments, 136 by other weapons, and 260 by personal attack (battery, strangulation, etc.). The FBI's Uniform Crime Report says that, among all murders nationwide, the weapon use rates were as follows: guns 63.4%, knives 13.1%, other weapons 16.8%, personal attack 6.7%. Is there evidence that violence against women involves different weapons than other violent attacks in the United States?
- Fruit flies.** Offspring of certain fruit flies may have yellow or ebony bodies and normal wings or short wings. Genetic theory predicts that these traits will appear in the ratio 9:3:3:1 (9 yellow, normal: 3 yellow, short: 3 ebony, normal: 1 ebony, short). A researcher checks 100 such flies and finds the distribution of the traits to be 59, 20, 11, and 10, respectively.

  - Are the results this researcher observed consistent with the theoretical distribution predicted by the genetic model?
  - If the researcher had examined 200 flies and counted exactly twice as many in each category—118, 40, 22, 20—what conclusion would he have reached?
  - Why is there a discrepancy between the two conclusions?
- Pi.** Many people know the mathematical constant  $\pi$  is approximately 3.14. But that's not exact. To be more precise, here are 20 decimal places: 3.14159265358979323846. Still not exact, though. In fact, the actual value is irrational, a decimal that goes on forever without any repeating pattern. But notice that there are no 0's and only one 7 in the 20 decimal places above. Does that pattern persist, or do all the digits show up with equal frequency? The table shows the number of times each digit appears in the first million digits. Test the hypothesis that the digits 0 through 9 are uniformly distributed in the decimal representation of  $\pi$ .
- Hurricane frequencies.** The National Hurricane Center provides data that list the numbers of large (category 3, 4, or 5) hurricanes that have struck the United States, by decade since 1851 ([http://www.nhc.noaa.gov/Deadliest\\_Costliest.shtml](http://www.nhc.noaa.gov/Deadliest_Costliest.shtml)). The data are on the next page.

The first million digits of  $\pi$

Digit	Count
0	99,959
1	99,758
2	100,026
3	100,229
4	100,230
5	100,359
6	99,548
7	99,800
8	99,985
9	100,106

Decade	Count	Decade	Count
1851–1860	6	1931–1940	8
1861–1870	1	1941–1950	10
1871–1880	7	1951–1960	9
1881–1890	5	1961–1970	6
1891–1900	8	1971–1980	4
1901–1910	4	1981–1990	4
1911–1920	7	1991–2000	5
1921–1930	5	2001–2006	7

Recently, there's been some concern that perhaps the number of large hurricanes has been increasing. The natural null hypothesis would be that the frequency of such hurricanes has remained constant.

- With 96 large hurricanes observed over the 16 periods, what are the expected value(s) for each cell?
- What kind of chi-square test would be appropriate?
- State the null and alternative hypotheses.
- How many degrees of freedom are there?
- The value of  $\chi^2$  is 12.67. What's the P-value?
- State your conclusion.
- Look again at the definition of the last "decade". Does that alter your conclusion at all?

- T 12. Lottery numbers.** The fairness of the South African lottery was recently challenged by one of the country's political parties. The lottery publishes historical statistics at its Website (<http://www.nationallottery.co.za/lotto/statistics.aspx>). Here is a table of the number of times each of the 49 numbers has been drawn in the main lottery and as the "bonus ball" number as of June 2007:

Number	Count	Bonus	Number	Count	Bonus
1	81	14	26	78	12
2	91	16	27	83	16
3	78	14	28	76	7
4	77	12	29	76	12
5	67	16	30	99	16
6	87	12	31	78	10
7	88	15	32	73	15
8	90	16	33	81	14
9	80	9	34	81	13
10	77	19	35	77	15
11	84	12	36	73	8
12	68	14	37	64	17
13	79	9	38	70	11
14	90	12	39	67	14
15	82	9	40	75	13
16	103	15	41	84	11
17	78	14	42	79	8
18	85	14	43	74	14
19	67	18	44	87	14
20	90	13	45	82	19
21	77	13	46	91	10
22	78	17	47	86	16
23	90	14	48	88	21
24	80	8	49	76	13
25	65	11			

We wonder if all the numbers are equally likely to be the "bonus ball".

- What kind of test should we perform?
  - There are 655 bonus ball observations. What are the appropriate expected value(s) for the test?
  - State the null and alternative hypotheses.
  - How many degrees of freedom are there?
  - The value of  $\chi^2$  is 34.5. What's the P-value?
  - State your conclusion.
- 13. Childbirth, part 1.** There is some concern that if a woman has an epidural to reduce pain during childbirth, the drug can get into the baby's bloodstream, making the baby sleepier and less willing to breastfeed. In December 2006, the *International Breastfeeding Journal* published results of a study conducted at Sydney University. Researchers followed up on 1178 births, noting whether the mother had an epidural and whether the baby was still nursing after 6 months. Here are their results:

		Epidural?		Total
		Yes	No	
Breastfeeding @ 6 months?	Yes	206	498	704
	No	190	284	474
Total		396	782	1178

- What kind of test would be appropriate?
  - State the null and alternative hypotheses.
- 14. Does your doctor know?** A survey<sup>7</sup> of articles from the *New England Journal of Medicine (NEJM)* classified them according to the principal statistics methods used. The articles recorded were all non-editorial articles appearing during the indicated years. Let's just look at whether these articles used statistics at all.

	Publication Year			Total
	1978–79	1989	2004–05	
No stats	90	14	40	144
Stats	242	101	271	614
Total	332	115	311	758

Has there been a change in the use of Statistics?

- What kind of test would be appropriate?
  - State the null and alternative hypotheses.
- 15. Childbirth, part 2.** In Exercise 13, the table shows results of a study investigating whether aftereffects of epidurals administered during childbirth might interfere with successful breastfeeding. We're planning to do a chi-square test.
- How many degrees of freedom are there?
  - The smallest expected count will be in the epidural/no breastfeeding cell. What is it?
  - Check the assumptions and conditions for inference.

<sup>7</sup>Suzanne S. Switzer and Nicholas J. Horton, "What Your Doctor Should Know about Statistics (but Perhaps Doesn't)" *Chance*, 20:1, 2007.

16. **Does your doctor know? (part 2).** The table in Exercise 14 shows whether *NEJM* medical articles during various time periods included statistics or not. We're planning to do a chi-square test.
- How many degrees of freedom are there?
  - The smallest expected count will be in the 1989/No cell. What is it?
  - Check the assumptions and conditions for inference.
17. **Childbirth, part 3.** In Exercises 13 and 15, we've begun to examine the possible impact of epidurals on successful breastfeeding.
- Calculate the component of chi-square for the epidural/no breastfeeding cell.
  - For this test,  $\chi^2 = 14.87$ . What's the P-value?
  - State your conclusion.
18. **Does your doctor know? (part 3).** In Exercises 14 and 16, we've begun to examine whether the use of statistics in *NEJM* medical articles has changed over time.
- Calculate the component of chi-square for the 1989/No cell.
  - For this test,  $\chi^2 = 25.28$ . What's the P-value?
  - State your conclusion.
19. **Childbirth, part 4.** In Exercises 13, 15, and 17, we've tested a hypothesis about the impact of epidurals on successful breastfeeding. The table shows the test's residuals.

		Epidural?	
		Yes	No
Breastfeeding at 6 months?	Yes	-1.99	1.42
	No	2.43	-1.73

- Show how the residual for the epidural/no breastfeeding cell was calculated.
  - What can you conclude from the standardized residuals?
20. **Does your doctor know? (part 4).** In Exercises 14, 16, and 18, we've tested a hypothesis about whether the use of statistics in *NEJM* medical articles has changed over time. The table shows the test's residuals.

	1978-79	1989	2004-05
No stats	3.39	-1.68	-2.48
Stats	-1.64	0.81	1.20

- Show how the residual for the 1989/No cell was calculated.
  - What can you conclude from the patterns in the standardized residuals?
21. **Childbirth, part 5.** In Exercises 13, 15, 17, and 19, we've looked at a study examining epidurals as one factor that might inhibit successful breastfeeding of newborn babies. Suppose a broader study included several additional issues, including whether the mother drank alcohol, whether this was a first child, and whether the parents occasionally supplemented breastfeeding with bottled formula. Why would it not be appropriate to use

chi-square methods on the  $2 \times 8$  table with yes/no columns for each potential factor?

22. **Does your doctor know? (part 5).** In Exercises 14, 16, 18, and 20, we considered data on articles in the *NEJM*. The original study listed 23 different Statistics methods. (The list read: *t*-tests, contingency tables, linear regression, . . .) Why would it not be appropriate to use a chi-square test on the  $23 \times 3$  table with a row for each method?

- T 23. **Titanic.** Here is a table we first saw in Chapter 3 showing who survived the sinking of the *Titanic* based on whether they were crew members, or passengers booked in first-, second-, or third-class staterooms:

	Crew	First	Second	Third	Total
Alive	212	202	118	178	710
Dead	673	123	167	528	1491
Total	885	325	285	706	2201

- If we draw an individual at random, what's the probability that we will draw a member of the crew?
- What's the probability of randomly selecting a third-class passenger who survived?
- What's the probability of a randomly selected passenger surviving, given that the passenger was a first-class passenger?
- If someone's chances of surviving were the same regardless of their status on the ship, how many members of the crew would you expect to have lived?
- State the null and alternative hypotheses.
- Give the degrees of freedom for the test.
- The chi-square value for the table is 187.8, and the corresponding P-value is barely greater than 0. State your conclusions about the hypotheses.

- T 24. **NYPD and sex discrimination.** The table below shows the rank attained by male and female officers in the New York City Police Department (NYPD). Do these data indicate that men and women are equitably represented at all levels of the department?

		Male	Female
		Rank	
Officer	21,900	4,281	
Detective	4,058	806	
Sergeant	3,898	415	
Lieutenant	1,333	89	
Captain	359	12	
Higher ranks	218	10	

- What's the probability that a person selected at random from the NYPD is a female?
- What's the probability that a person selected at random from the NYPD is a detective?
- Assuming no bias in promotions, how many female detectives would you expect the NYPD to have?



- d) To see if there is evidence of differences in ranks attained by males and females, will you test goodness-of-fit, homogeneity, or independence?
- e) State the hypotheses.
- f) Test the conditions.
- g) How many degrees of freedom are there?
- h) The chi-square value for the table is 290.1 and the P-value is less than 0.0001. State your conclusion about the hypotheses.

25. **Titanic again.** Examine and comment on this table of the standardized residuals for the chi-square test you looked at in Exercise 23.

	Crew	First	Second	Third
Alive	-4.35	9.49	2.72	-3.30
Dead	3.00	-6.55	-1.88	2.27

26. **NYPD again.** Examine and comment on this table of the standardized residuals for the chi-square test you looked at in Exercise 24.

	Male	Female
Officer	-2.34	5.57
Detective	-1.18	2.80
Sergeant	3.84	-9.14
Lieutenant	3.58	-8.52
Captain	2.46	-5.86
Higher ranks	1.74	-4.14

- T 27. Cranberry juice.** It's common folk wisdom that drinking cranberry juice can help prevent urinary tract infections in women. In 2001 the *British Medical Journal* reported the results of a Finnish study in which three groups of 50 women were monitored for these infections over 6 months. One group drank cranberry juice daily, another group drank a lactobacillus drink, and the third drank neither of those beverages, serving as a control group. In the control group, 18 women developed at least one infection, compared to 20 of those who consumed the lactobacillus drink and only 8 of those who drank cranberry juice. Does this study provide supporting evidence for the value of cranberry juice in warding off urinary tract infections?
- a) Is this a survey, a retrospective study, a prospective study, or an experiment? Explain.
  - b) Will you test goodness-of-fit, homogeneity, or independence?
  - c) State the hypotheses.
  - d) Test the conditions.
  - e) How many degrees of freedom are there?
  - f) Find  $\chi^2$  and the P-value.
  - g) State your conclusion.
  - h) If you concluded that the groups are not the same, analyze the differences using the standardized residuals of your calculations.

**T 28. Cars.** A random survey of autos parked in the student lot and the staff lot at a large university classified the brands by country of origin, as seen in the table. Are there differences in the national origins of cars driven by students and staff?

		Driver	
		Student	Staff
Origin	American	107	105
	European	33	12
	Asian	55	47

- a) Is this a test of independence or homogeneity?
  - b) Write appropriate hypotheses.
  - c) Check the necessary assumptions and conditions.
  - d) Find the P-value of your test.
  - e) State your conclusion and analysis.
- T 29. Montana.** A poll conducted by the University of Montana classified respondents by whether they were male or female and political party, as shown in the table. We wonder if there is evidence of an association between being male or female and party affiliation.

	Democrat	Republican	Independent
Male	36	45	24
Female	48	33	16

- a) Is this a test of homogeneity or independence?
  - b) Write an appropriate hypothesis.
  - c) Are the conditions for inference satisfied?
  - d) Find the P-value for your test.
  - e) State a complete conclusion.
- T 30. Fish diet.** Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer. ("Fatty Fish Consumption and Risk of Prostate Cancer," *Lancet*, June 2001)

Fish Consumption	Total Subjects	Prostate Cancers
Never/seldom	124	14
Small part of diet	2621	201
Moderate part	2978	209
Large part	549	42

- a) Is this a survey, a retrospective study, a prospective study, or an experiment? Explain.
- b) Is this a test of homogeneity or independence?
- c) Do you see evidence of an association between the amount of fish in a man's diet and his risk of developing prostate cancer?
- d) Does this study prove that eating fish does not prevent prostate cancer? Explain.

- T 31. Montana revisited.** The poll described in Exercise 29 also investigated the respondents' party affiliations based on what area of the state they lived in. Test an appropriate hypothesis about this table and state your conclusions.

	Democrat	Republican	Independent
West	39	17	12
Northeast	15	30	12
Southeast	30	31	16

- T 32. Working parents.** In July 1991 and again in April 2001, the Gallup Poll asked random samples of 1015 adults about their opinions on working parents. The table summarizes responses to the question "Considering the needs of both parents and children, which of the following do you see as the ideal family in today's society?"

	1991	2001
Both work full time	142	131
One works full time, other part time	274	244
One works, other works at home	152	173
One works, other stays home for kids	396	416
No opinion	51	51

- a) Is this a survey, a retrospective study, a prospective study, or an experiment? Explain.
- b) Will you test goodness-of-fit, homogeneity, or independence?
- c) Based on these results, do you think there was a change in people's attitudes during the 10 years between these polls?
- 33. Grades.** Two different professors teach an introductory Statistics course. The table shows the distribution of final grades they reported. We wonder whether one of these professors is an "easier" grader.

	Prof. Alpha	Prof. Beta
A	3	9
B	11	12
C	14	8
D	9	2
F	3	1

- a) Will you test goodness-of-fit, homogeneity, or independence?
- b) Write appropriate null hypotheses.
- c) Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate.
- T 34. Full moon.** Some people believe that a full moon elicits unusual behavior in people. The table shows the number of arrests made in a small town during weeks of six full

moons and six other randomly selected weeks in the same year. We wonder if there is evidence of a difference in the types of illegal activity that take place.

	Full Moon	Not Full
Violent (murder, assault, rape, etc.)	2	3
Property (burglary, vandalism, etc.)	17	21
Drugs/Alcohol	27	19
Domestic abuse	11	14
Other offenses	9	6

- a) Will you test goodness-of-fit, homogeneity, or independence?
- b) Write appropriate null hypotheses.
- c) Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate.
- 35. Grades again.** In some situations where the expected cell counts are too small, as in the case of the grades given by Professors Alpha and Beta in Exercise 33, we can complete an analysis anyway. We can often proceed after combining cells in some way that makes sense and also produces a table in which the conditions are satisfied. Here we create a new table displaying the same data, but calling D's and F's "Below C":

	Prof. Alpha	Prof. Beta
A	3	9
B	11	12
C	14	8
Below C	12	3

- a) Find the expected counts for each cell in this new table, and explain why a chi-square procedure is now appropriate.
- b) With this change in the table, what has happened to the number of degrees of freedom?
- c) Test your hypothesis about the two professors, and state an appropriate conclusion.
- T 36. Full moon, next phase.** In Exercise 34 you found that the expected cell counts failed to satisfy the conditions for inference.
- a) Find a sensible way to combine some cells that will make the expected counts acceptable.
- b) Test a hypothesis about the full moon and state your conclusion.
- T 37. Racial steering.** A subtle form of racial discrimination in housing is "racial steering." Racial steering occurs when real estate agents show prospective buyers only homes in neighborhoods already dominated by that family's race. This violates the Fair Housing Act of 1968. According to an article in *Chance* magazine (Vol. 14, no. 2 [2001]), tenants at a large apartment complex recently filed a lawsuit alleging racial steering. The complex is divided into two parts: Section A and Section B.

The plaintiffs claimed that white potential renters were steered to Section A, while African-Americans were steered to Section B. The table displays the data that were presented in court to show the locations of recently rented apartments. Do you think there is evidence of racial steering?

New Renters			
	White	Black	Total
Section A	87	8	95
Section B	83	34	117
Total	170	42	212

- T 38. Titanic, redux.** Newspaper headlines at the time, and traditional wisdom in the succeeding decades, have held that women and children escaped the *Titanic* in greater proportions than men. Here's a table with the relevant data. Do you think that survival was independent of whether the person was male or female? Explain.

	Female	Male	Total
Alive	343	367	710
Dead	127	1364	1491
Total	470	1731	2201

- 39. Steering revisited.** You could have checked the data in Exercise 37 for evidence of racial steering using two-proportion  $z$  procedures.
- Find the  $z$ -value for this approach, and show that when you square your  $z$ -value, you get the value of  $\chi^2$  you calculated in Exercise 37.
  - Show that the resulting  $P$ -values are the same.
- T 40. Survival on the Titanic, one more time.** In Exercise 38 you could have checked for a difference in the chances of survival for men and women using two-proportion  $z$  procedures.
- Find the  $z$ -value for this approach.
  - Show that the square of your calculated value of  $z$  is the value of  $\chi^2$  you calculated in Exercise 38.
  - Show that the resulting  $P$ -values are the same.
- T 41. Pregnancies.** Most pregnancies result in live births, but some end in miscarriages or stillbirths. A June 2001 National Vital Statistics Report examined those outcomes in the United States during 1997, broken down by the age of the mother. The table shows counts consistent with that report. Is there evidence that the distribution of outcomes is not the same for these age groups?

Age of Mother	Live Births	Fetal Losses
	Under 20	49
20–29	201	41
30–34	88	21
35 or over	49	21

- T 42. Education by age.** Use the survey results in the table to investigate differences in education level attained among different age groups in the United States.

Education Level	Age Group				
	25–34	35–44	45–54	55–64	≥ 65
Not HS grad	27	50	52	71	101
HS	82	19	88	83	59
1–3 years college	43	56	26	20	20
≥ 4 years college	48	75	34	26	20



### JUST CHECKING Answers

- We need to know how well beetles can survive 6 hours in a Plexiglas<sup>®</sup> box so that we have a baseline to compare the treatments.
- There's no difference in survival rate in the three groups.
- $(2 - 1)(3 - 1) = 2 \text{ df}$
- 50
- 2
- The mean value for a  $\chi^2$  with 2 df is 2, so 10 seems pretty large. The  $P$ -value is probably small.
- This is a test of homogeneity. The clue is that the question asks whether the distributions are alike.
- This is a test of goodness-of-fit. We want to test the model of equal assignment to all lots against what actually happened.
- This is a test of independence. We have responses on two variables for the same individuals.

# Inferences for Regression



<b>WHO</b>	250 male subjects
<b>WHAT</b>	Body fat and waist size
<b>UNITS</b>	% Body fat and inches
<b>WHEN</b>	1990s
<b>WHERE</b>	United States
<b>WHY</b>	Scientific research

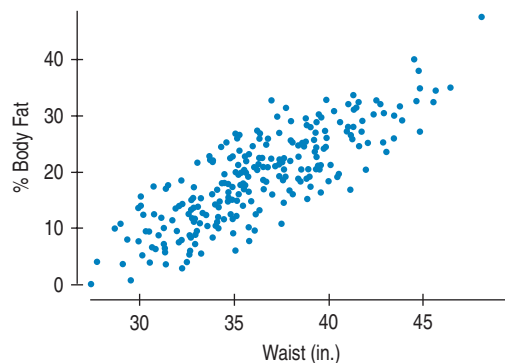
**T**hree percent of a man's body is essential fat. (For a woman, the percentage is closer to 12.5%.) As the name implies, essential fat is necessary for a normal, healthy body. Fat is stored in small amounts throughout your body. Too much body fat, however, can be dangerous to your health. For men between 18 and 39 years old, a healthy percent body fat ranges from 8% to 19%. (For women of the same age, it's 21% to 32%.)

Measuring body fat can be tedious and expensive. The "standard reference" measurement is by dual-energy X-ray absorptiometry (DEXA), which involves two low-dose X-ray generators and takes from 10 to 20 minutes.

How close can we get to a useable prediction of body fat from easily measurable variables such as *Height*, *Weight*, or *Waist* size? Here's a scatterplot of *%Body Fat* plotted against *Waist* size for a sample of 250 males of various ages.

**FIGURE 27.1**

Percent Body Fat vs. Waist size for 250 men of various ages. The scatterplot shows a strong, positive, linear relationship.



Back in Chapter 8 we modeled relationships like this by fitting a least squares line. The plot is clearly straight, so we can find that line. The equation of the least squares line for these data is

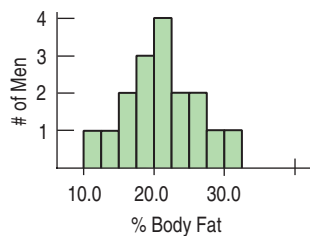
$$\widehat{\%Body\ Fat} = -42.7 + 1.7\ Waist.$$

The slope says that, on average, *%Body Fat* is greater by 1.7 percent for each additional inch around the waist.

How useful is this model? When we fit linear models before, we used them to describe the relationship between the variables and we interpreted the slope and intercept as descriptions of the data. Now we'd like to know what the regression model can tell us beyond the 250 men in this study. To do that, we'll want to make confidence intervals and test hypotheses about the slope and intercept of the regression line.

## The Population and the Sample

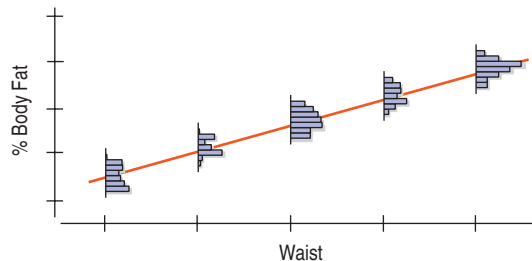
When we found a confidence interval for a mean, we could imagine a single, true underlying value for the mean. When we tested whether two means or two proportions were equal, we imagined a true underlying difference. But what does it mean to do inference for regression? We know better than to think that even if we knew every population value, the data would line up perfectly on a straight line. After all, even in our sample, not all men who have 38-inch waists have the same %Body Fat. In fact, there's a whole distribution of %Body Fat for these men:



**FIGURE 27.2**

The distribution of %Body Fat for men with a Waist size of 38 inches is unimodal and symmetric.

This is true at each *Waist* size. In fact, we could depict the distribution of %Body Fat at different *Waist* sizes like this:



**FIGURE 27.3**

There's a distribution of %Body Fat for each value of *Waist* size. We'd like the means of these distributions to line up.

But we want to *model* the relationship between %Body Fat and *Waist* size for all men. To do that, we imagine an idealized regression line. The model assumes that the means of the distributions of %Body Fat for each *Waist* size fall along the line, even though the individuals are scattered around it. We know that this model is not a perfect description of how the variables are associated, but it may be useful for predicting %Body Fat and for understanding how it's related to *Waist* size.

If only we had all the values in the population, we could find the slope and intercept of this *idealized regression line* explicitly by using least squares. Following our usual conventions, we write the idealized line with Greek letters and consider the coefficients (the slope and intercept) to be *parameters*:  $\beta_0$  is the intercept and  $\beta_1$  is the slope. Corresponding to our fitted line of  $\hat{y} = b_0 + b_1x$ , we write

$$\mu_y = \beta_0 + \beta_1x.$$

Why  $\mu_y$  instead of  $\hat{y}$ ? Because this is a model. There is a distribution of %Body Fat for each *Waist* size. The model places the *means* of the distributions of %Body Fat for each *Waist* size on the same straight line.

### NOTATION ALERT:

This time we used up only one Greek letter for two things. Lower-case Greek  $\beta$  (beta) is the natural choice to correspond to the  $b$ 's in the regression equation. We used  $\beta$  before for the probability of a Type II error, but there's little chance of confusion here.

Of course, not all the individual  $y$ 's are at these means. (In fact, the line will miss most—and quite possibly all—of the plotted points.) Some individuals lie above and some below the line, so, like all models, this one makes **errors**. Lots of them. In fact, one at each point. These errors are random and, of course, can be positive or negative. They are model errors, so we use a Greek letter and denote them by  $\varepsilon$ .

When we put the errors into the equation, we can account for each individual  $y$ :

$$y = \beta_0 + \beta_1x + \varepsilon.$$

This equation is now true for each data point (since there is an  $\varepsilon$  to soak up the deviation), so the model gives a value of  $y$  for any value of  $x$ .

For the body fat data, an idealized model such as this provides a summary of the relationship between *%Body Fat* and *Waist* size. Like all models, it simplifies the real situation. We know there is more to predicting body fat than waist size alone. But the advantage of a model is that the simplification might help us to think about the situation and assess how well *%Body Fat* can be predicted from simpler measurements.

We estimate the  $\beta$ 's by finding a regression line,  $\hat{y} = b_0 + b_1x$ , as we did in Chapter 8. The residuals,  $e = y - \hat{y}$ , are the sample-based versions of the errors,  $\varepsilon$ . We'll use them to help us assess the regression model.

We know that least squares regression will give reasonable estimates of the parameters of this model from a random sample of data. Our challenge is to account for our uncertainty in how well they do. For that, we need to make some assumptions about the model and the errors.

## Assumptions and Conditions

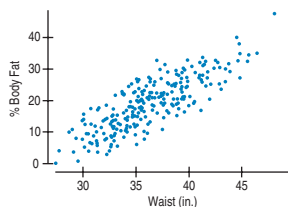
**AS** **Activity: Conditions for Regression Inference.** View an illustrated discussion of the conditions for regression inference.

Back in Chapter 8 when we fit lines to data, we needed to check only the Straight Enough Condition. Now, when we want to make inferences about the coefficients of the line, we'll have to make more assumptions. Fortunately, we can check conditions to help us judge whether these assumptions are reasonable for our data. And as we've done before, we'll make some checks *after* we find the regression equation.

Also, we need to be careful about the order in which we check conditions. If our initial assumptions are not true, it makes no sense to check the later ones. So now we number the assumptions to keep them in order.

### Check the scatterplot.

The shape must be linear or we can't use linear regression at all.



### 1. LINEARITY ASSUMPTION

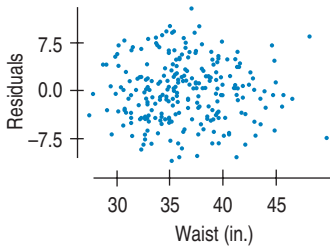
If the true relationship is far from linear and we use a straight line to fit the data, our entire analysis will be useless, so we always check this first.

The **Straight Enough Condition** is satisfied if a scatterplot looks straight. It's generally not a good idea to draw a line through the scatterplot when checking. That can fool your eyes into seeing the plot as more straight. Sometimes it's easier to see violations of the Straight Enough Condition by looking at a scatterplot of the residuals against  $x$  or against the predicted values,  $\hat{y}$ . That plot will have a horizontal direction and should have no pattern if the condition is satisfied.

If the scatterplot is straight enough, we can go on to some assumptions about the errors. If not, stop here, or consider re-expressing the data (see Chapter 10) to make the scatterplot more nearly linear. For the *%Body Fat* data, the scatterplot is beautifully linear. Of course, the data must be quantitative for this to make sense. Check the **Quantitative Data Condition**.

**Check the residuals plot (1).**

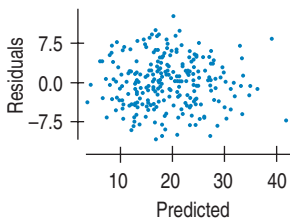
The residuals should appear to be randomly scattered.

**FIGURE 27.4**

The residuals show only random scatter when plotted against Waist size.

**Check the residuals plot (2).**

The vertical spread of the residuals should be roughly the same everywhere.

**FIGURE 27.5**

A scatterplot of residuals against predicted values can help check for plot thickening. Note that this plot looks identical to the plot of residuals against Waist size. For a regression of one response variable on one predictor, these plots differ only in the labels on the  $x$ -axis.

**2. INDEPENDENCE ASSUMPTION**

**Independence Assumption:** The errors in the true underlying regression model (the  $\varepsilon$ 's) must be mutually independent. As usual, there's no way to be sure that the Independence Assumption is true.

Usually when we care about inference for the regression parameters, it's because we think our regression model might apply to a larger population. In such cases, we can check a **Randomization Condition** that the individuals are a representative sample from that population.

We can also check displays of the regression residuals for evidence of patterns, trends, or clumping, any of which would suggest a failure of independence. In the special case when the  $x$ -variable is related to time, a common violation of the Independence Assumption is for the errors to be correlated. (The error our model makes today may be similar to the one it made for yesterday.) This violation can be checked by plotting the residuals against the  $x$ -variable and looking for patterns.

The %Body Fat data were collected on a sample of men taken to be representative. The subjects were not related in any way, so we can be pretty sure that their measurements are independent. The residuals plot shows no pattern.

**3. EQUAL VARIANCE ASSUMPTION**

The variability of  $y$  should be about the same for all values of  $x$ . In Chapter 8 we looked at the standard deviation of the residuals ( $s_e$ ) to measure the size of the scatter. Now we'll need this standard deviation to build confidence intervals and test hypotheses. The standard deviation of the residuals is the building block for the standard errors of all the regression parameters. But it makes sense only if the scatter of the residuals is the same everywhere. In effect, the standard deviation of the residuals "pools" information across all of the individual distributions at each  $x$ -value, and pooled estimates are appropriate only when they combine information for groups with the same variance.

Practically, what we can check is the **Does the Plot Thicken? Condition**. A scatterplot of  $y$  against  $x$  offers a visual check. Fortunately, we've already made one. Make sure the spread around the line is nearly constant. Be alert for a "fan" shape or other tendency for the variation to grow or shrink in one part of the scatterplot. Often it is better to look at the residuals plotted against the predicted values,  $\hat{y}$ . With the slope of the line removed, it's easier to see patterns left behind. For the body fat data, the spread of %Body Fat around the line is remarkably constant across Waist sizes from 30 inches to about 45 inches.

If the plot is straight enough, the data are independent, and the plot doesn't thicken, you can now move on to the final assumption.

**4. NORMAL POPULATION ASSUMPTION**

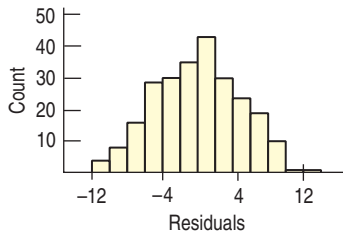
We assume the errors around the idealized regression line at each value of  $x$  follow a Normal model. We need this assumption so that we can use a Student's  $t$ -model for inference.

As we have at other times when we've used Student's  $t$ , we'll settle for the residuals satisfying the **Nearly Normal Condition** and the **Outlier Condition**. Look at a histogram or Normal probability plot of the residuals.<sup>1</sup>

<sup>1</sup> This is why we have to check the conditions in order. We have to check that the residuals are independent and that the variation is the same for all  $x$ 's so that we can lump all the residuals together for a single check of the Nearly Normal Condition.

**Check a histogram of the residuals.**

The distribution of the residuals should be unimodal and symmetric.

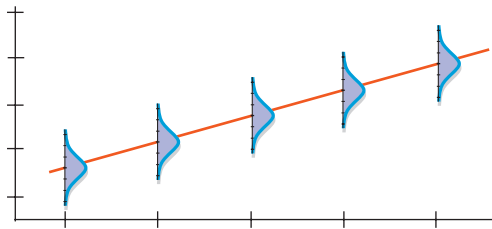


**FIGURE 27.6**

A histogram of the residuals is one way to check whether they are nearly Normal. Alternatively, we can look at a Normal probability plot.

The histogram of residuals in the %Body Fat regression certainly looks nearly Normal. As we have noted before, the Normality Assumption becomes less important as the sample size grows, because the model is about means and the Central Limit Theorem takes over.

If all four assumptions were true, the idealized regression model would look like this:



**FIGURE 27.7**

The regression model has a distribution of  $y$ -values for each  $x$ -value. These distributions follow a normal model with means lined up along the line and with the same standard deviations.

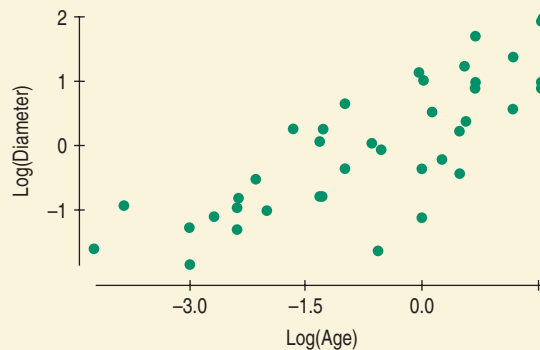
At each value of  $x$  there is a distribution of  $y$ -values that follows a Normal model, and each of these Normal models is centered on the line and has the same standard deviation. Of course, we don't expect the assumptions to be exactly true, and we know that all models are wrong, but the linear model is often close enough to be very useful.

**FOR EXAMPLE**

**Checking assumptions and conditions**

Look at the moon with binoculars or a telescope, and you'll see craters formed by thousands of impacts. The earth, being larger, has been hit even more often. Meteor Crater in Arizona was the first recognized impact crater and was identified as such only in the 1920s. With the help of satellite images, more and more craters have been identified; now more than 180 are known. These, of course, are only a small sample of all the impacts the earth has experienced: Only 29% of earth's surface is land, and many craters have been covered or eroded away. Astronomers have recognized a roughly 35 million-year cycle in the frequency of cratering, although the cause of this cycle is not fully understood. Here's a scatterplot of the known impact craters from the most recent 35 million years.<sup>2</sup> We've taken logs of both age (in millions of years ago) and diameter (km) to make the relationship simpler. (See Chapter 10.)

- WHO** 39 impact craters
- WHAT** Diameter and age
- UNITS** km and millions of years ago
- WHEN** Past 35 million years
- WHERE** Worldwide
- WHY** Scientific research



**Question:** Are the assumptions and conditions satisfied for fitting a linear regression model to these data?

- ✓ **Linearity Assumption:** The scatterplot satisfies the Straight Enough Condition.
- ✓ **Independence Assumption:** Sizes of impact craters are likely to be generally independent.

(continued)

<sup>2</sup> Data, pictures, and much more information at the Earth Impact Database found at <http://www.unb.ca>.

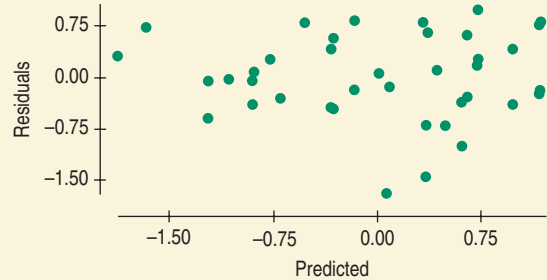


For Example (continued)

✓ **Randomization Condition:** These are the only known craters, and may differ from others that have disappeared or not yet been found. I'll need to be careful not to generalize my conclusions too broadly.

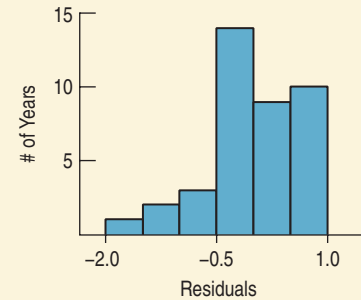
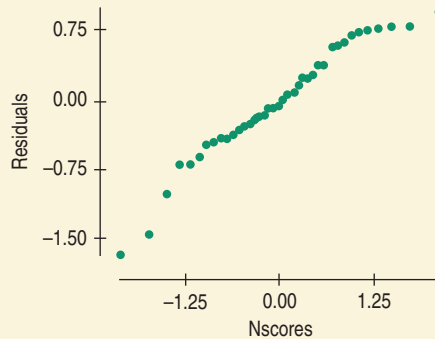
✓ **Does the Plot Thicken? Condition:** After fitting a linear model, I find the residuals shown.

Two points seem to give the impression that the residuals may be more variable for higher predicted values than for lower ones, but this doesn't seem to be a serious violation of the Equal Variance Assumption.



✓ **Nearly Normal Condition:** A Normal probability plot suggests a bit of skewness in the distribution of residuals, and the histogram confirms that.

There are no violations severe enough to stop my regression analysis, but I'll be cautious about my conclusions.



## Which Come First: the Conditions or the Residuals?

*“Truth will emerge more readily from error than from confusion.”*

—Francis Bacon  
(1561–1626)

In regression, there's a little catch. The best way to check many of the conditions is with the residuals, but we get the residuals only *after* we compute the regression. Before we compute the regression, however, we should check at least one of the conditions.

So we work in this order:

1. Make a scatterplot of the data to check the Straight Enough Condition. (If the relationship is curved, try re-expressing the data. Or stop.)
2. If the data are straight enough, fit a regression and find the residuals,  $e$ , and predicted values,  $\hat{y}$ .
3. Make a scatterplot of the residuals against  $x$  or the predicted values. This plot should have no pattern. Check in particular for any bend (which would suggest that the data weren't all that straight after all), for any thickening (or thinning), and, of course, for any outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions—one with and one without the outliers.)
4. If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not independent.
5. If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
6. If all the conditions seem to be reasonably satisfied, go ahead with inference.

## STEP-BY-STEP EXAMPLE

## Regression Inference

If our data can jump through all these hoops, we're ready to do regression inference. Let's see how much more we can learn about body fat and waist size from a regression model.

**Questions:** What is the relationship between %Body Fat and Waist size in men?

What model best predicts body fat from waist size, and how well does it do the job?

THINK

**Plan** Specify the question of interest.

Name the variables and report the W's.

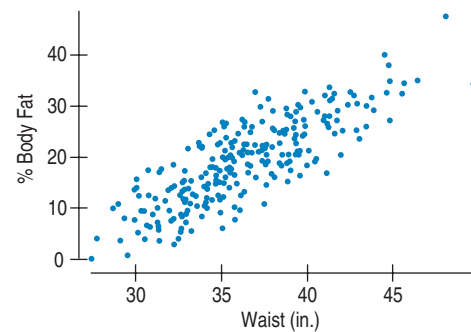
Identify the parameters you want to estimate.

**Model** Think about the assumptions and check the conditions.

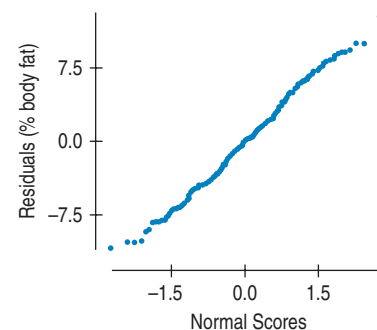
Make pictures. For regression inference, you'll need a scatterplot, a residuals plot, and either a histogram or a Normal probability plot of the residuals.

(We've seen plots of the residuals already. See Figures 27.5 and 27.6.)

I have quantitative body measurements on 250 adult males from the BYU Human Performance Research Center. I want to understand the relationship between %Body Fat and Waist size.



- ✓ **Straight Enough Condition:** There's no obvious bend in the original scatterplot of the data or in the plot of residuals against predicted values.
- ✓ **Independence Assumption:** These data are not collected over time, and there's no reason to think that the %Body Fat of one man influences the %Body Fat of another.
- ✓ **Does the Plot Thicken? Condition:** Neither the original scatterplot nor the residual scatterplot shows any changes in the spread about the line.
- ✓ **Nearly Normal Condition, Outlier Condition:** A histogram of the residuals is unimodal and symmetric. The Normal probability plot of the residuals is quite straight, indicating that the Normal model is reasonable for the errors.



Choose your method.

Under these conditions a **regression model** is appropriate.



**Mechanics** Let's just "push the button" and see what the regression looks like.

The formula for the regression equation can be found in Chapter 8, and the standard error formulas will be shown a bit later, but regressions are almost always computed with a computer program or calculator.

Write the regression equation.

Here's the computer output for this regression:

Dependent variable is: %BF

R-squared = 67.8%

s = 4.713 with 250 - 2 = 248 degrees of freedom

Variable	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	-42.734	2.717	-15.7	<0.0001
Waist	1.70	0.0743	22.9	<0.0001

The estimated regression equation is

$$\widehat{\%Body\ Fat} = -42.73 + 1.70\ Waist.$$



**Conclusion** Interpret your results in context.

**More Interpretation** We haven't worked it out in detail yet, but the output gives us numbers labeled as *t*-statistics and corresponding P-values, and we have a general idea of what those mean.

(Now it's time to learn more about regression inference so we can figure out what the rest of the output means.)

The  $R^2$  for the regression is 67.8%. Waist size seems to account for about 2/3 of the %Body Fat variation in men. The slope of the regression says that %Body Fat increases by about 1.7 percentage points per inch of Waist size, on average.

The standard error of 0.07 for the slope is much smaller than the slope itself, so it looks like the estimate is reasonably precise. And there are a couple of *t*-ratios and P-values given. Because the P-values are small, it appears that some null hypotheses can be rejected.

## Intuition About Regression Inference

**A S** **Simulation:** Simulate the Sampling Distribution of a Regression Slope. Draw samples repeatedly to see for yourself how slope can vary from sample to sample. This simulation experiment lets you build up a histogram to see the sampling distribution.

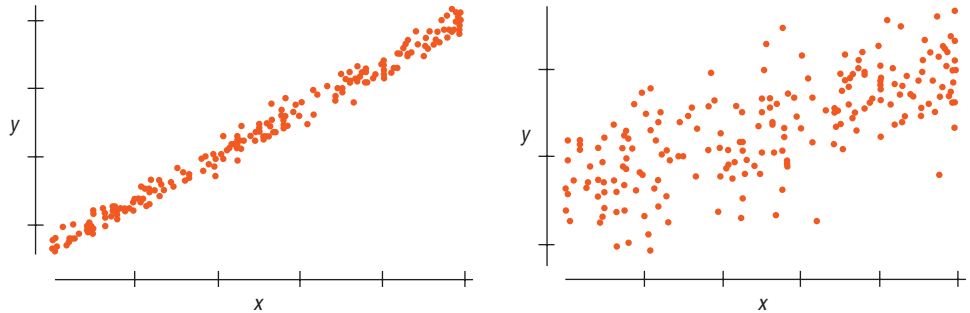
Wait a minute! We've just pulled a fast one. We've pushed the "regression button" on our computer or calculator but haven't discussed where the standard errors for the slope or intercept come from. We know that if we had collected similar data on a different random sample of men, the slope and intercept would be different. Each sample would have produced its own regression line, with slightly different  $b_0$ 's and  $b_1$ 's. This sample-to-sample variation is what generates the sampling distributions for the coefficients.

There's only one regression model; each sample regression is trying to estimate the same parameters,  $\beta_0$  and  $\beta_1$ . We expect any sample to produce a  $b_1$  whose expected value is the true slope,  $\beta_1$ . What about its standard deviation? What aspects of the data affect how much the slope (and intercept) vary from sample to sample?

- ▶ **Spread around the line.** Here are two situations in which we might do regression. Which situation would yield the more consistent slope? That is, if we were to sample over and over from the two underlying populations that these samples come from and compute all the slopes, which group of slopes would vary less?

**FIGURE 27.8**

Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from its underlying population?



***n* - 2?**

For standard deviation (in Chapter 4), we divided by  $n - 1$  because we didn't know the true mean and had to estimate it. Now it's later in the course and there's even more we don't know. Here we don't know *two* things: the slope and the intercept. If we knew them both, we'd divide by  $n$  and have  $n$  degrees of freedom. When we estimate both, however, we adjust by subtracting 2, so we divide by  $n - 2$  and (as we will see soon) have 2 fewer degrees of freedom.

Clearly, data like those in the left plot give more consistent slopes.

Less scatter around the line means the slope will be more consistent from sample to sample. The spread around the line is measured with the **residual standard deviation,  $s_e$** . You can always find  $s_e$  in the regression output, often just labeled  $s$ . You're probably not going to calculate the residual standard deviation by hand. As we noted when we first saw this formula in Chapter 8, it looks a lot like the standard deviation of  $y$ , only now subtracting the predicted values rather than the mean and dividing by  $n - 2$  instead of  $n - 1$ :

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

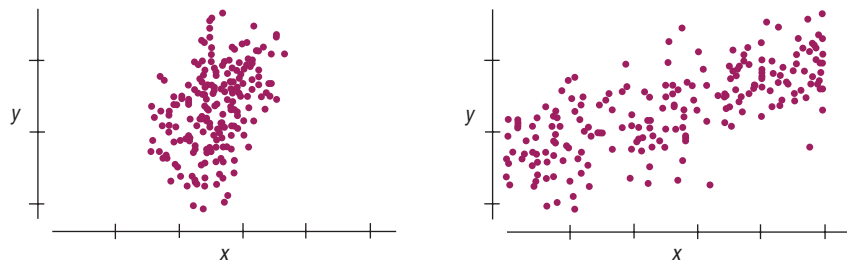
The less scatter around the line, the smaller the residual standard deviation and the stronger the relationship between  $x$  and  $y$ .

Some people prefer to assess the strength of a regression by looking at  $s_e$  rather than  $R^2$ . After all,  $s_e$  has the same units as  $y$ , and because it's the standard deviation of the errors around the line, it tells you how close the data are to our model. By contrast,  $R^2$  is the proportion of the variation of  $y$  accounted for by  $x$ . We say, why not look at both?

- ▶ **Spread of the  $x$ 's:** Here are two more situations. Which of these would yield more consistent slopes?

**FIGURE 27.9**

Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from the underlying population?

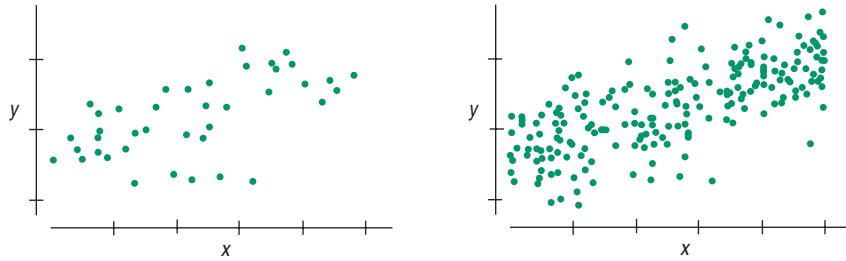


A plot like the one on the right has a broader range of  $x$ -values, so it gives a more stable base for the slope. We'd expect the slopes of samples from situations like that to vary less from sample to sample. A large standard deviation of  $x$ ,  $s_x$ , provides a more stable regression.

► **Sample size.** Here we go again. What about these two?

FIGURE 27.10

Which of these scatterplots shows a situation that would give the more consistent regression slope estimate if we were to sample repeatedly from the underlying population?



It shouldn't be a surprise that having a larger sample size,  $n$ , gives more consistent estimates from sample to sample.

## Standard Error for the Slope

Three aspects of the scatterplot, then, affect the standard error of the regression slope:

- Spread around the line:  $s_e$
- Spread of  $x$  values:  $s_x$
- Sample size:  $n$

These are in fact the *only* things that affect the standard error of the slope. Although you'll probably never have to calculate it by hand, the formula for the standard error is

$$SE(b_1) = \frac{s_e}{\sqrt{n-1} s_x}.$$

The error standard deviation,  $s_e$ , is in the *numerator*, since spread around the line *increases* the slope's standard error. The denominator has both a sample size term  $\sqrt{n-1}$  and  $s_x$ , because increasing either of these *decreases* the slope's standard error.

We know the  $b_1$ 's vary from sample to sample. As you'd expect, their sampling distribution model is centered at  $\beta_1$ , the slope of the idealized regression line. Now we can estimate its standard deviation with  $SE(b_1)$ . What about its shape? Here the Central Limit Theorem and "Wild Bill" Gosset come to the rescue again. When we standardize the slopes by subtracting the model mean and dividing by their standard error, we get a Student's  $t$ -model, this time with  $n-2$  degrees of freedom:

$$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}.$$

**AS** **Activity: Regression Slope Standard Error.** See how  $SE(b_1)$  is constructed and where the values used in the formula are found in the regression output table.

**AS** **Simulation:  $x$ -Variance and Slope Variance.** You don't have to just imagine how the variability of the slope depends on the spread of the  $x$ 's.

### NOTATION ALERT:

Don't confuse the standard deviation of the residuals,  $s_e$ , with the standard error of the slope,  $SE(b_1)$ . The first measures the scatter around the line, and the second tells us how reliably we can estimate the slope.

### A SAMPLING DISTRIBUTION FOR REGRESSION SLOPES

When the conditions are met, the standardized estimated regression slope,

$$t = \frac{b_1 - \beta_1}{SE(b_1)},$$

follows a Student's  $t$ -model with  $n-2$  degrees of freedom. We estimate the standard error with

$$SE(b_1) = \frac{s_e}{\sqrt{n-1} s_x}, \text{ where } s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

$n$  is the number of data values, and  $s_x$  is the ordinary standard deviation of the  $x$ -values.

## FOR EXAMPLE

### Finding standard errors

**Recap:** Recent terrestrial impact craters seem to show a relationship between age and size that is linear when re-expressed using logarithms (see Chapter 10).

Here are summary statistics and regression output.

Variable	Count	Mean	StdDev
LogAge	39	-0.656310	1.57682
LogDiam	39	0.012600	1.04104

Dependent variable is: LogDiam

R-squared = 63.6%

s = 0.6362 with 39 - 2 = 37 degrees of freedom

Variable	Coefficient	Se(coeff)	t-ratio	P-value
Intercept	0.358262	0.1106	3.24	0.0025
LogAge	0.526674	0.0655	8.05	≤ 0.0001

**Questions:** How are the standard error of the slope and the  $t$ -ratio for the slope calculated? (And aren't you glad the software does this for you?)

$$SE(b_1) = \frac{s_e}{\sqrt{n-1} \times s_x} = \frac{0.6362}{\sqrt{39-1} \times 1.57682} = 0.0655$$

$$\text{Assuming no linear association } (\beta_1 = 0), t_{37} = \frac{b_1 - \beta_1}{SE(b_1)} = \frac{0.526674 - 0}{0.0655} = 8.05$$

## What About the Intercept?

The same reasoning applies for the intercept. We could write

$$\frac{b_0 - \beta_0}{SE(b_0)} \sim t_{n-2}$$

and use it to construct confidence intervals and test hypotheses, but often the value of the intercept isn't something we care about. The intercept usually isn't interesting. Most hypothesis tests and confidence intervals for regression are about the slope.

## Regression Inference

### TI-*inspire*

**Regression Inference.** How big must a slope be in order to be considered statistically significant? See for yourself by exploring the natural sample-to-sample variability in slopes.

Now that we have the standard error of the slope and its sampling distribution, we can test a hypothesis about it and make confidence intervals. The usual null hypothesis about the slope is that it's equal to 0. Why? Well, a slope of zero would say that  $y$  doesn't tend to change linearly when  $x$  changes—in other words, that there is no linear association between the two variables. If the slope were zero, there wouldn't be much left of our regression equation.

So a null hypothesis of a zero slope questions the entire claim of a linear relationship between the two variables—and often that's just what we want to know. In fact, every software package or calculator that does regression simply assumes that you want to test the null hypothesis that the slope is really zero.

**What if the Slope Were 0?**

If  $b_1 = 0$ , our prediction is  $\hat{y} = b_0 + 0x$ . The equation collapses to just  $\hat{y} = b_0$ . Now  $x$  is nowhere in sight, so  $y$  doesn't depend on  $x$  at all.

And  $b_0$  would turn out to be  $\bar{y}$ . Why? We know that  $b_0 = \bar{y} - b_1\bar{x}$ , but when  $b_1 = 0$ , that becomes simply  $b_0 = \bar{y}$ . It turns out, then, that when the slope is 0, the equation is just  $\hat{y} = \bar{y}$ ; at every value of  $x$ , we always predict the mean value for  $y$ .

To test  $H_0: \beta_1 = 0$ , we find

$$t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$$

This is just like every  $t$ -test we've seen: a difference between the statistic and its hypothesized value, divided by its standard error.

For our body fat data, the computer found the slope (1.7), its standard error (0.0743), and the ratio of the two:  $\frac{1.7 - 0}{0.0743} = 22.9$  (see p. 656). Nearly 23 standard errors from the hypothesized value certainly seems big. The P-value ( $<0.0001$ ) confirms that a  $t$ -ratio this large would be very unlikely to occur if the true slope were zero.

Maybe the standard null hypothesis isn't all that interesting here. Did you have any doubts that %Body Fat is related to Waist size? A more sensible use of these same values might be to make a confidence interval for the slope instead.

We can build a confidence interval in the usual way, as an estimate plus or minus a margin of error. As always, the margin of error is just the product of the standard error and a critical value. Here the critical value comes from the  $t$ -distribution with  $n - 2$  degrees of freedom, so a 95% confidence interval for  $\beta$  is

$$b_1 \pm t_{n-2}^* \times SE(b_1)$$

For the body fat data,  $t_{248}^* = 1.970$ , so that comes to  $1.7 \pm 1.97 \times 0.074$ , or an interval from 1.55 to 1.85 %Body Fat per inch of Waist size.

**FOR EXAMPLE**

**Interpreting a regression model**

**Recap:** On a log scale, there seems to be a linear relationship between the diameter and the age of recent terrestrial impact craters. We have regression output from statistics software:

Dependent variable is: LogDiam  
 R-squared = 63.6%  
 s = 0.6362 with 39 - 2 = 37 degrees of freedom

**Questions:** What's the regression model, and what can it tell us?

Variable	Coefficient	Se(coeff)	t-ratio	P-value
Intercept	0.358262	0.1106	3.24	0.0025
LogAge	0.526674	0.0655	8.05	$\leq 0.0001$

For terrestrial impact craters younger than 35 million years, the logarithm of Diameter grows linearly with the logarithm of Age:  $\log \text{Diam} = 0.358 + 0.527 \log \text{Age}$ . The P-value for each coefficient's  $t$ -statistic is very small, so I'm quite confident that neither coefficient is zero. Based on my model, I conclude that, on average, the older a crater is, the larger it tends to be. This model accounts for 63.6% of the variation in  $\log \text{Diam}$ .

Although it is possible that impacts (and their craters) are getting smaller, it is more likely that I'm seeing the effects of age on craters. Small craters are probably more likely to erode or become buried or otherwise be difficult to find as they age. Larger craters may survive the huge expanses of geologic time more successfully.



**JUST CHECKING**

Researchers in Food Science studied how big people's mouths tend to be. They measured mouth volume by pouring water into the mouths of subjects who lay on their backs. Unless this is your idea of a good time, it would be helpful to have a model to estimate mouth volume more simply. Fortunately, mouth volume is related to height. (Mouth volume is measured in cubic centimeters and height in meters.)

The data were checked and deemed suitable for regression. Take a look at the computer output.

1. What does the  $t$ -ratio of 3.27 tell us about this relationship? How does the P-value help our understanding?
2. Would you say that measuring a person's height could reliably be used as a substitute for the wetter method of determining how big a person's mouth is? What numbers in the output helped you reach that conclusion?
3. What does the value of  $s_e$  add to this discussion?

Summary of	Mouth Volume			
Mean	60.2704			
StdDev	16.8777			
Dependent variable is:	Mouth Volume			
R-squared =	15.3%			
$s = 15.66$ with	$61 - 2 = 59$	degrees of freedom		
<b>Variable</b>	<b>Coefficient</b>	<b>SE(coeff)</b>	<b>t-ratio</b>	<b>P-value</b>
Intercept	-44.7113	32.16	-1.39	0.1697
Height	61.3787	18.77	3.27	0.0018

## Another Example



**AS** **Activity: A Hypothesis Test for the Regression Slope.**  
View an animated discussion of testing the standard null hypothesis for slope.

Every spring, Nenana, Alaska, hosts a contest in which participants try to guess the exact minute that a wooden tripod placed on the frozen Tanana River will fall through the breaking ice. The contest started in 1917 as a diversion for railroad engineers, with a jackpot of \$800 for the closest guess. It has grown into an event in which hundreds of thousands of entrants enter their guesses on the Internet<sup>3</sup> and vie for as much as \$300,000.

Because so much money and interest depends on the time of breakup, it has been recorded to the nearest minute with great accuracy ever since 1917. And because a standard measure of breakup has been used throughout this time, the data are consistent. An article in *Science*<sup>4</sup> used the data to investigate global warming—whether greenhouse gasses and other human actions have been making the planet warmer. Others might just want to make a good prediction of next year's breakup time.

Of course, we can't use regression to tell the *causes* of any change. But we can estimate the *rate* of change (if any) and use it to make better predictions.

Here are some of the data:

<b>WHO</b>	Years
<b>WHAT</b>	Year, day, and hour of ice breakup
<b>UNITS</b>	$x$ is in years since 1900. $y$ is in days after midnight Dec. 31.
<b>WHEN</b>	1917–present
<b>WHERE</b>	Nenana, Alaska
<b>WHY</b>	Wagering, but proposed to look at global warming

Year (since 1900)	Breakup Date (days after Jan. 1)	Year (since 1900)	Breakup Date (days after Jan. 1)
17	119.4792	30	127.7938
18	130.3979	31	129.3910
19	122.6063	32	121.4271
20	131.4479	33	127.8125
21	130.2792	34	119.5882
22	131.5556	35	134.5639
23	128.0833	36	120.5403
24	131.6319	37	131.8361
25	126.7722	38	125.8431
26	115.6688	39	118.5597
27	131.2375	40	110.6437
28	126.6840	41	122.0764
29	124.6535	:	:

<sup>3</sup> <http://www.nenanaaiceclassic.com>

<sup>4</sup> "Climate Change in Nontraditional Data Sets." *Science* 294 [26 October 2001]: 811.



## STEP-BY-STEP EXAMPLE

A Regression Slope  $t$ -Test

The slope of the regression gives the change in Nenana ice breakup date per year.

**Questions:** Is there sufficient evidence to claim that ice breakup times are changing?  
If so, how rapid is the change?

THINK

**Plan** State what you want to know.

Identify the *parameter* you wish to estimate. Here our parameter is the slope.

Identify the variables and review the  $W$ 's.

**Hypotheses** Write your null and alternative hypotheses.

**Model** Think about the assumptions and check the conditions.

Make pictures. Because the scatterplot seems straight enough, we can find and plot the residuals.

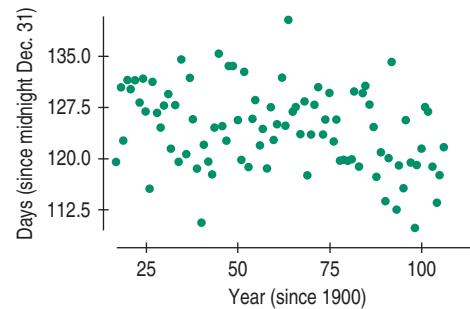
Usually, we check for suggestions that the Independence Assumption fails by plotting the residuals against the predicted values. Patterns and clusters in that plot raise our suspicions. But when the data are measured over time, it is always a good idea to plot residuals against time to look for trends and oscillations.

I wonder whether the date of ice breakup in Nenana has changed over time. The slope of that change might indicate climate change. I have the date of ice breakup annually since 1917, recorded as the number of days and fractions of a day until the ice breakup.

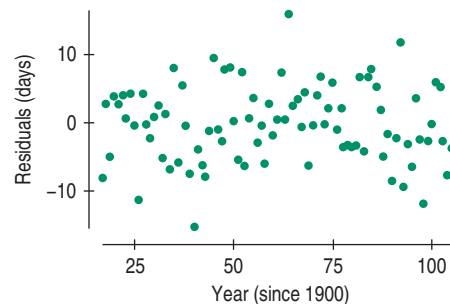
$H_0$ : There is no change in the date of ice breakup:  $\beta_1 = 0$

$H_A$ : Yes, there is:  $\beta_1 \neq 0$

✓ **Straight Enough Condition:** I have quantitative data with no obvious bend in the scatterplot.



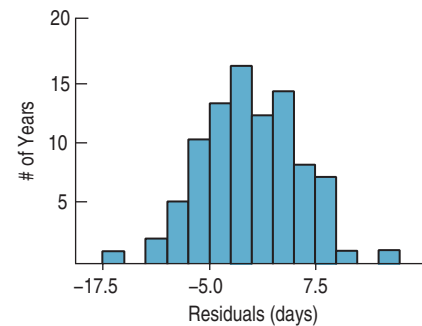
✓ **Independence Assumption:** These data are a time series, which raises my suspicions that they may not be independent. To check, here's a plot of the residuals against time, the  $x$ -variable of the regression:



I see a hint that the data oscillate up and down, which suggests some failure of independence, but not so strongly that I can't

proceed with the analysis. These data are not a random sample, so I'm reluctant to extend my conclusions beyond this river and these years.

- ✓ **Does the Plot Thicken? Condition:** The residuals plot shows no obvious trends in the spread.
- ✓ **Nearly Normal Condition, Outlier Condition:** A histogram of the residuals is unimodal and symmetric.



Under these conditions, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with  $(n - 2) = 89$  degrees of freedom.

I'll do a **regression slope *t*-test**.

State the sampling distribution model.

Choose your method.

**SHOW**

**Mechanics** The regression equation can be found from the formulas in Chapter 8, but regressions are almost always found from a computer program or calculator.

The P-values given in the regression output table are from the Student's *t*-distribution on  $(n - 2) = 89$  degrees of freedom. They are appropriate for two-sided alternatives.

Here's the computer output for this regression:

Dependent variable is: Breakup Date

R-squared = 11.3%

$s = 5.673$  with  $91 - 2 = 89$  degrees of freedom

Variable	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	128.950	1.525	84.6	<0.0001
Year Since 1900	-0.07606	0.0226	-3.36	0.0012

The estimated regression equation is

$$\widehat{\text{Date}} = 128.95 - 0.076 \text{ YearSince1900}.$$

**TELL**

**Conclusion** Link the P-value to your decision and state your conclusion in the proper context.

The P-value of 0.0012 means that the association we see in the data is unlikely to have occurred by chance. I reject the null hypothesis, and conclude that there is strong evidence that, on average, the ice breakup is occurring earlier each year. But the oscillation pattern in the residuals raises concerns.



### Create a confidence interval for the true slope

A 95% confidence interval for  $\beta_1$  is

$$b_1 \pm t_{89}^* \times SE(b_1)$$

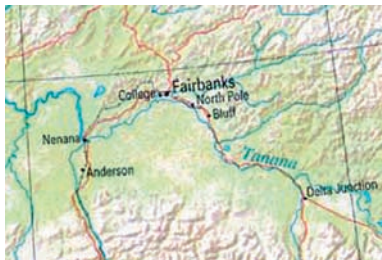
$$-0.076 \pm (1.987)(0.0226)$$

or  $(-0.12, -0.03)$  days per year.



**Interpret the interval** Simply rejecting the standard null hypothesis doesn't guarantee that the size of the effect is large enough to be important. Whether we want to know the breakup time to the nearest minute or are interested in global warming, a change measured in hours each year is big enough to be interesting.

I am 95% confident that the ice has been breaking up, on average, between 0.03 days (about 40 minutes) and 0.12 days (about 3 hours) earlier each year since 1900.



**But is it global warming?** So the ice is breaking up earlier. Temperatures are higher. Must be global warming, right?

Maybe.

An article challenging the original analysis of the Nenana data proposed a possible confounding variable. It noted that the city of Fairbanks is upstream from Nenana and suggested that the growth of Fairbanks could have warmed the river. So maybe it's not global warming.

Or maybe global warming is a lurking variable, leading more people to move to a now balmy Fairbanks and also leading to generally earlier ice breakup in Nenana.

Or maybe there's some other variable or combination of variables at work. We can't set up an experiment, so we may never really know.

Only one thing is for sure. When you try to explain an association by claiming cause and effect, you're bound to be on thin ice.<sup>5</sup>

### TI Tips

°F	Min
44	142.7
46	142.1
47	143.4
50	143.6
51	144.0
52	143.4
54	142.4
55	143.1
57	143.7
60	143.4
65	143.4

### Doing regression inference

The TI will easily do almost everything you need for inference for regression: scatterplots, residual plots, histograms of residuals, and  $t$ -tests and confidence intervals for the slope of the regression line. OK, it won't tell you  $SE(b)$ , but it will give you enough information to easily figure it out for yourself. Not bad.

As an example we'll use data from *Chance* magazine (Vol. 12, No. 4, 1999), giving times and temperatures for 11 of the top performances in women's marathons during the 1990s. Let's examine the influence of temperature on the performance of elite runners in marathons.

<sup>5</sup> How *do* scientists sort out such messy situations? Even though they can't conduct an experiment, they *can* look for replications elsewhere. A number of studies of ice on other bodies of water have also shown earlier ice breakup times in recent years. That suggests they need an explanation that's more comprehensive than just Fairbanks and Nenana.

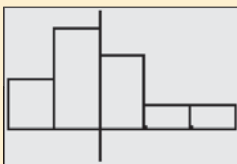
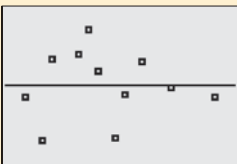


```
LinRegTTest
Xlist:L1
Ylist:L2
Freq:1
8 & P-Test <0 >0
RegEQ:Y1
Calculate
```

```
LinRegTTest
y=a+bx
a≠0 and b≠0
t=1.135800072
P=.2853799627
df=9
↓a=141.4824942
```

```
LinRegTTest
y=a+bx
a≠0 and b≠0
↑b=.032517321
s=.5680158733
r²=.1253679846
r=.354073417
```

```
LinRegTInt
y=a+bx
(-.0322, .09728)
b=.032517321
df=9
s=.5680158733
↓a=141.4824942
```



### Test a Hypothesis About the Association

- Enter the temperatures (nearest degree Fahrenheit) in **L1** and the runners' times (nearest tenth of a minute) in **L2**.
- Check the scatterplot. It's not obviously nonlinear, so go ahead.
- Under **STAT TESTS** choose **LinRegTTest**.
- Specify the two data lists (with **Freq:1**).
- Choose the two-tailed option. (We are interested in whether higher temperatures enhance or interfere with a runner's performance.)
- Tell it to store the regression equation in **Y1** (**VARS**, **Y-VARS**, **Function** . . . remember?), then **Calculate**.

The TI creates so much information you have to scroll down to see it all! Look what's there.

- The calculated value of **t** and the **P**-value.
- The coefficients of the regression equation, **a** and **b**.
- The value of **s**, our sample estimate of the common standard deviation of errors around the true line.
- The values of **r<sup>2</sup>** and **r**.

Wait, where's  $SE(b)$ ? It's not there. No problem—if you need it, you can figure it out. Remember that the  $t$ -value is  $b$  divided by  $SE(b)$ . So  $SE(b)$  must be  $b$  divided by  $t$ . Here  $SE(b) = 0.0325 \div 1.1358 = 0.0286$ .

### Create a Confidence Interval for the Slope

- Back to **STAT TEST**; this time you want **LinRegTInt**.
- The specifications for the data lists and the regression equation remain what you entered for the hypothesis test.
- Choose a confidence level, say 95%, and **Calculate**.

### Checking Conditions

Beware!!! Before you try to interpret any of this, you must check the conditions to see if inference for regression is allowed.

- We already looked at the scatterplot; it was reasonably linear.
- To create the residuals plot, set up another scatterplot with **RESID** (from **LIST NAMES**) as your **Ylist**. OK, it looks fairly random.
- The residuals plot may show a slight hint of diminishing scatter, but with so few data values it's not very clear.
- The histogram of the residuals is unimodal and roughly symmetric.

### What Does It All Mean?

Because the conditions check out okay, we can try to summarize what we have learned. With a P-value over 28%, it's quite possible that any perceived relationship could be just sampling error. The confidence interval suggests the slope could be positive or negative, so it's possible that as temperatures increase, women marathoners may run faster—or slower. Based on these 11 races there appears to be little evidence of a linear association between temperature and women's performances in the marathon.

## \* Standard Errors for Predicted Values

Once we have a useful regression, how can we indulge our natural desire to predict, without being irresponsible? We know how to compute predicted values of  $y$  for any value of  $x$ . We first did that in Chapter 8. This predicted value would be our best estimate, but it's still just an informed guess.

Now, however, we have standard errors. We can use those to construct a confidence interval for the predictions and to report our uncertainty honestly.

From our model of %Body Fat and Waist size, we might want to use Waist size to get a reasonable estimate of %Body Fat. A confidence interval can tell us how precise that prediction will be. The precision depends on the question we ask, however, and there are two questions: Do we want to know the mean %Body Fat for all men with a Waist size of, say, 38 inches? Or do we want to estimate the %Body Fat for a particular man with a 38-inch Waist without making him climb onto the X-ray table?

What's the difference between the two questions? The predicted %Body Fat is the same, but one question leads to an answer much more precise than the other. We can predict the mean %Body Fat for all men whose Waist size is 38 inches with a lot more precision than we can predict the %Body Fat of a particular individual whose Waist size happens to be 38 inches. Both are interesting questions.

We start with the same prediction in both cases. We are predicting the value for a new individual, one that was not part of the original data set. To emphasize this, we'll call his  $x$ -value " $x$  sub new" and write it  $x_\nu$ .<sup>6</sup> Here,  $x_\nu$  is 38 inches. The regression equation predicts %Body Fat as  $\hat{y}_\nu = b_0 + b_1x_\nu$ .

Now that we have the predicted value, we construct both intervals around this same number. Both intervals take the form

$$\hat{y}_\nu \pm t_{n-2}^* \times SE.$$

Even the  $t^*$  value is the same for both. It's the critical value (from Table T or technology) for  $n - 2$  degrees of freedom and the specified confidence level. The intervals differ because they have different standard errors. Our choice of ruler depends on which interval we want.

The standard errors for prediction depend on the same kinds of things as the coefficients' standard errors. If there is more spread around the line, we'll be less certain when we try to predict the response. Of course, if we're less certain of the slope, we'll be less certain of our prediction. If we have more data, our estimate will be more precise. And there's one more piece: If we're farther from the center of our data, our prediction will be less precise. This last factor is new but makes intuitive sense: It's a lot easier to predict a data point near the middle of the data set than far from the center.

Each of these factors contributes uncertainty—that is, variability—to the estimate. Because the factors are independent of each other, we can add their variances to find the total variability. The resulting formula for the standard error of the predicted mean value explicitly takes into account each of the factors:

$$SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}.$$

Individual values vary more than means, so the standard error for a single predicted value has to be larger than the standard error for the mean. In fact, the standard error of a single predicted value has an *extra* source of variability: the variation of individuals around the predicted mean. That appears as the extra variance term,  $s_e^2$ , at the end under the square root:

$$SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}.$$

For the Nenana Ice Classic, someone who planned to place a bet would want to predict this year's breakup time. By contrast, scientists studying global warming are likely to be more interested in the mean breakup time. Unfortunately if you want to gamble, the variability is greater for predicting for a single year.

<sup>6</sup> Yes, this is a bilingual pun. The Greek letter  $\nu$  is called "nu." Don't blame me; my co-author suggested this.

Keep in mind this distinction between the two kinds of confidence intervals: The narrower interval is a **confidence interval for the predicted mean value at  $x_v$** , and the wider interval is a **prediction interval for an individual with that  $x$ -value**.

## FOR EXAMPLE

### \*Finding confidence intervals for predicted values

Let's use our analysis to create confidence intervals for predictions about %Body Fat. From the data and the regression output we know:

$$n = 250 \quad \bar{x} = 36.3 \quad s_e = 4.713 \quad SE(b_1) = 0.074$$

**Question 1:** What's a 95% confidence interval for the mean %Body Fat for all men with 38-inch waists?

For  $x_v = 38$  the regression model predicts

$$\hat{y}_v = -42.7 + 1.7(38) = 21.9\%.$$

The standard error is

$$SE(\hat{\mu}_v) = \sqrt{0.074^2(38 - 36.3)^2 + \frac{4.713^2}{250}} = 0.32\%.$$

With  $250 - 2 = 248$  df, for 95% confidence  $t^* = 1.97$ .

Putting it all together, the 95% confidence interval is:  $21.9\% \pm 1.97(0.32)$

$$21.9\% \pm 0.63\%, \text{ or } (21.27, 22.53)$$

I'm 95% confident that the mean body fat level for all men with 38-inch waists is between 21.3% and 22.5% body fat.

**Question 2:** What's a 95% prediction interval for the %Body Fat of an individual man with a 38-inch waist?

The standard error is

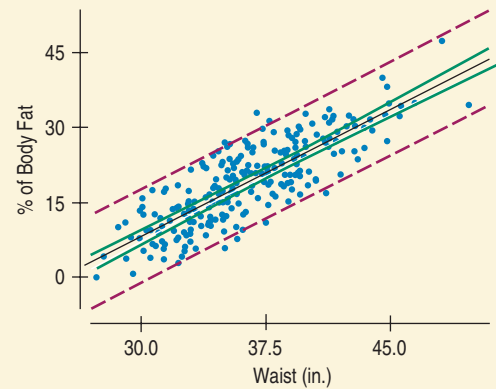
$$SE(\hat{y}_v) = \sqrt{0.074^2(38 - 36.3)^2 + \frac{4.713^2}{250} + 4.713^2} = 4.72\%.$$

The prediction interval is:  $21.9\% \pm 1.97(4.72)$

$$21.9\% \pm 9.3\%, \text{ or } (12.6, 31.2)$$

I'm 95% confident that a randomly selected man with a 38-inch waist will have between 12.6% and 31.2% body fat.

Notice how much wider this interval is than the first one. As we've known since Chapter 18, the mean is such less variable than a randomly selected individual value.



**FIGURE 27.11**

A scatterplot of %Body Fat vs. Waist size with a least squares regression line. The solid green lines near the regression line show the extent of the 95% confidence intervals for mean %Body Fat at each Waist size. The dashed red lines show the prediction intervals. Most of the points are contained within the prediction intervals, but not within the confidence intervals.

**\*MATH BOX**

So where do those messy formulas for standard errors of predicted values come from? They're based on many of the ideas we've studied so far. Start with regression, add random variables, then throw in the Pythagorean Theorem, the Central Limit Theorem, and a dose of algebra. Mix well. . .

We begin our quest with an equation of the regression line. Usually we write the line in the form  $\hat{y} = b_0 + b_1x$ . Mathematicians call that the "slope-intercept" form; in your algebra class you wrote it as  $y = mx + b$ . In that algebra class you also learned another way to write equations of lines. When you know that a line with slope  $m$  passes through the point  $(x_1, y_1)$ , the "point-slope" form of its equation is  $y - y_1 = m(x - x_1)$ .

We know the regression line passes through the mean-mean point  $(\bar{x}, \bar{y})$  with slope  $b_1$ , so we can write its equation in point-slope form as  $\hat{y} - \bar{y} = b_1(x - \bar{x})$ . Solving for  $\hat{y}$  yields  $\hat{y} = b_1(x - \bar{x}) + \bar{y}$ . This equation predicts the mean  $y$ -value for a specific  $x_v$ :

$$\hat{\mu}_y = b_1(x_v - \bar{x}) + \bar{y}.$$

To create a confidence interval for the mean value we need to measure the variability in this prediction:

$$\text{Var}(\hat{\mu}_y) = \text{Var}(b_1(x_v - \bar{x}) + \bar{y}).$$

We now call on the Pythagorean Theorem of Statistics once more: the slope,  $b_1$ , and mean,  $\bar{y}$ , should be independent, so their variances add:

$$\text{Var}(\hat{\mu}_y) = \text{Var}(b_1(x_v - \bar{x})) + \text{Var}(\bar{y}).$$

The horizontal distance from our specific  $x$ -value to the mean,  $x_v - \bar{x}$ , is a constant:

$$\text{Var}(\hat{\mu}_y) = (\text{Var}(b_1))(x_v - \bar{x})^2 + \text{Var}(\bar{y}).$$

Let's write that equation in terms of standard deviations:

$$\text{SD}(\hat{\mu}_y) = \sqrt{(\text{SD}^2(b_1))(x_v - \bar{x})^2 + \text{SD}^2(\bar{y})}.$$

Because we'll need to estimate these standard deviations using samples statistics, we're really dealing with standard errors:

$$\text{SE}(\hat{\mu}_y) = \sqrt{(\text{SE}^2(b_1))(x_v - \bar{x})^2 + \text{SE}^2(\bar{y})}.$$

The Central Limit Theorem tells us that the standard deviation of  $\bar{y}$  is  $\frac{\sigma}{\sqrt{n}}$ . Here we'll estimate  $\sigma$  using  $s_e$ , which describes the variability in how far the line we drew through our sample mean may lie above or below the true mean:

$$\begin{aligned} \text{SE}(\hat{\mu}_y) &= \sqrt{(\text{SE}^2(b_1))(x_v - \bar{x})^2 + \left(\frac{s_e}{\sqrt{n}}\right)^2} \\ &= \sqrt{(\text{SE}^2(b_1))(x_v - \bar{x})^2 + \frac{s_e^2}{n}}. \end{aligned}$$

And there it is—the standard error we need to create a confidence interval for a predicted mean value.

When we try to predict an individual value of  $y$ , we must also worry about how far the true point may lie above or below the regression line. We represent that uncertainty by adding another term,  $e$ , to the original equation:

$$y = b_1(x_v - \bar{x}) + \bar{y} + e.$$

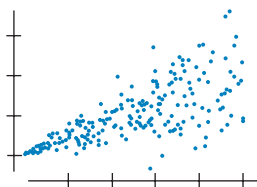
To make a long story short (and the equation a wee bit longer), that additional term simply adds one more standard error to the sum of the variances:

$$\text{SE}(\hat{y}) = \sqrt{(\text{SE}^2(b_1))(x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}.$$

## WHAT CAN GO WRONG?

In this chapter we've added inference to the regression explorations that we did in Chapters 8 and 9. Everything covered in those chapters that could go wrong with regression can still go wrong. It's probably a good time to review Chapter 9. Take your time; we'll wait.

With inference, we've put numbers on our estimates and predictions, but these numbers are only as good as the model. Here are the main things to watch out for:



▶ **Don't fit a linear regression to data that aren't straight.** This is the most fundamental assumption. If the relationship between  $x$  and  $y$  isn't approximately linear, there's no sense in fitting a straight line to it.

▶ **Watch out for the plot thickening.** The common part of confidence and prediction intervals is the estimate of the error standard deviation, the spread around the line. If it changes with  $x$ , the estimate won't make sense. Imagine making a prediction interval for these data.

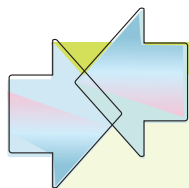
When  $x$  is small, we can predict  $y$  precisely, but as  $x$  gets larger, it's much harder to pin  $y$  down. Unfortunately, if the spread changes, the single value of  $s_e$  won't pick that up. The prediction interval will use the average spread around the line, with the result that we'll be too pessimistic about our precision for low  $x$ -values and too optimistic for high  $x$ -values. A re-expression of  $y$  is often a good fix for changing spread.

▶ **Make sure the errors are Normal.** When we make a prediction interval for an individual, the Central Limit Theorem can't come to our rescue. For us to believe the prediction interval, the errors must be from the Normal model. Check the histogram and Normal probability plot of the residuals to see if this assumption looks reasonable.

▶ **Watch out for extrapolation.** It's tempting to think that because we have prediction intervals, they'll take care of all our uncertainty so we don't have to worry about extrapolating. Wrong. The interval is only as good as the model. The uncertainty our intervals predict is correct only if our model is true. There's no way to adjust for wrong models. That's why it's always dangerous to predict for  $x$ -values that lie far from the center of the data.

▶ **Watch out for influential points and outliers.** We always have to be on the lookout for a few points that have undue influence on our estimated model—and regression is certainly no exception.

▶ **Watch out for one-tailed tests.** Because tests of hypotheses about regression coefficients are usually two-tailed, software packages report two-tailed P-values. If you are using software to conduct a one-tailed test about slope, you'll need to divide the reported P-value in half.



## CONNECTIONS

Regression inference is connected to almost everything we've done so far. Scatterplots are essential for checking linearity and whether the plot thickens. Histograms and normal probability plots come into play to check the Nearly Normal condition. And we're still thinking about the same attributes of the data in these plots as we were back in the first part of the book.

Regression inference is also connected to just about every inference method we have seen for measured data. The assumption that the spread of data about the line is constant is essentially the same as the assumption of equal variances required for the pooled- $t$  methods. Our use of all the residuals together to estimate their standard deviation is a form of pooling.



Inference for regression is closely related to inference for means, so your understanding of means transfers directly to your understanding of regression. Here's a table that displays the similarities:

	Means	Regression Slope
Parameter	$\mu$	$\beta_1$
Statistic	$\bar{y}$	$b_1$
Population spread estimate	$s_y = \sqrt{\frac{\sum(y - \bar{y})^2}{n - 1}}$	$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$
Standard error of the statistic	$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$	$SE(b_1) = \frac{s_e}{s_x \sqrt{n - 1}}$
Test statistic	$\frac{\bar{y} - \mu_0}{SE(\bar{y})} \sim t_{n-1}$	$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$
Margin of error	$ME = t_{n-1}^* \times SE(\bar{y})$	$ME = t_{n-2}^* \times SE(b_1)$

## WHAT HAVE WE LEARNED?



In Chapters 7, 8, and 9, we learned to examine the relationship between two quantitative variables in a scatterplot, to summarize its strength with correlation, and to fit linear relationships by least squares regression. And we saw that these methods are particularly powerful and effective for modeling, predicting, and understanding these relationships.

Now we have completed our study of inference methods by applying them to these regression models. We've found that the same methods we used for means—Student's  $t$ -models—work for regression in much the same way as they did for means. And we've seen that although this makes the mechanics familiar, there are new conditions to check and a need for care in describing the hypotheses we test and the confidence intervals we construct.

- ▶ We've learned that under certain assumptions, the sampling distribution for the slope of a regression line can be modeled by a Student's  $t$ -model with  $n - 2$  degrees of freedom.
- ▶ We've learned to check four conditions to verify those assumptions before we proceed with inference. We've learned the importance of checking these conditions in order, and we've seen that most of the checks can be made by graphing the data and the residuals with the methods we learned in Chapters 4, 5, and 8.
- ▶ We've learned to use the appropriate  $t$ -model to test a hypothesis about the slope. If the slope of our regression line is significantly different from zero, we have strong evidence that there is an association between the two variables.
- ▶ We've also learned to create and interpret a confidence interval for the true slope.
- ▶ And we've been reminded yet again never to mistake the presence of an association for proof of causation.

## Terms

Conditions for inference in regression (and checks for some of them)

- ▶ 651. **Straight Enough Condition** for linearity. (Check that the scatterplot of  $y$  against  $x$  has linear form and that the scatterplot of residuals against predicted values has no obvious pattern.)
- ▶ 652. **Independence Assumption.** (Think about the nature of the data. Check a residuals plot.)
- ▶ 652. **Does the Plot Thicken? Condition** for constant variance. (Check that the scatterplot shows consistent spread across the range of the  $x$ -variable, and that the residuals plot has constant variance, too. A common problem is increasing spread with increasing predicted values—the *plot thickens!*)

**Residual standard deviation**

- ▶ 652. **Nearly Normal Condition** for Normality of the residuals. (Check a histogram of the residuals.)

657. The spread of the data around the regression line is measured with the residual standard deviation,  $s_e$ :

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} = \sqrt{\frac{\sum e^2}{n - 2}}$$

 **$t$ -test for the regression slope**

658, 662. When the assumptions are satisfied, we can perform a test for the slope coefficient. We usually test the null hypothesis that the true value of the slope is zero against the alternative that it is not. A zero slope would indicate a complete absence of linear relationship between  $y$  and  $x$ .

To test  $H_0: \beta_1 = 0$ , we find

$$t = \frac{b_1 - 0}{SE(b_1)}$$

where

$$SE(b_1) = \frac{s_e}{\sqrt{n - 1} s_x}, \quad s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

$n$  is the number of cases, and  $s_x$  is the standard deviation of the  $x$ -values. We find the P-value from the Student's  $t$ -model with  $n - 2$  degrees of freedom.

**Confidence interval for the regression slope ( $\beta$ )**

660. When the assumptions are satisfied, we can find a confidence interval for the slope parameter from  $b_1 \pm t_{n-2}^* \times SE(b_1)$ . The critical value,  $t_{n-2}^*$ , depends on the confidence level specified and on Student's  $t$ -model with  $n - 2$  degrees of freedom.

**Skills****THINK**

- ▶ Understand that the “true” regression line does not fit the population data perfectly, but rather is an idealized summary of that data.
- ▶ Know how to examine your data and a scatterplot of  $y$  vs.  $x$  for violations of assumptions that would make inference for regression unwise or invalid.
- ▶ Know how to examine displays of the residuals from a regression to double-check that the conditions required for regression have been met. In particular, know how to judge linearity and constant variance from a scatterplot of residuals against predicted values. Know how to judge Normality from a histogram and Normal probability plot.
- ▶ Remember to be especially careful to check for failures of the Independence Assumption when working with data recorded over time. To search for patterns, examine scatterplots both of  $x$  against time and of the residuals against time.

**SHOW**

- ▶ Know how to test the standard hypothesis that the true regression slope is zero. Be able to state the null and alternative hypotheses. Know where to find the relevant numbers in standard computer regression output.
- ▶ Be able to find a confidence interval for the slope of a regression based on the values reported in a standard regression output table.

**TELL**

- ▶ Be able to summarize a regression in words. In particular, be able to state the meaning of the true regression slope, the standard error of the estimated slope, and the standard deviation of the errors.
- ▶ Be able to interpret the P-value of the  $t$ -statistic for the slope to test the standard null hypothesis.
- ▶ Be able to interpret a confidence interval for the slope of a regression.

## REGRESSION ANALYSIS ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables differ slightly from one package to another, but all are essentially the same. We've seen two examples of such tables already.

All packages offer analyses of the residuals. With some, you must request plots of the residuals as you request the regression. Others let you find the regression first and then analyze the residuals afterward. Either way, your analysis is not complete if you don't check the residuals with a histogram or Normal probability plot and a scatterplot of the residuals against  $x$  or the predicted values.

You should, of course, always look at the scatterplot of your two variables before computing a regression.

Regressions are almost always found with a computer or calculator. The calculations are too long to do conveniently by hand for data sets of any reasonable size. No matter how the regression is computed, the results are usually presented in a table that has a standard form. Here's a portion of a typical regression results table, along with annotations showing where the numbers come from:

**A S** **Activity: Regression on the Computer.** How fast is the universe expanding? And how old is it? A prominent astronomer used regression to astound the scientific community. Read the story, analyze the data, and interactively learn about each of the numbers in a typical computer regression output table.

Variable	Coefficient	SE(Coeff)	t-ratio	Prob
Constant	-42.7341	2.717	-15.7	$\leq 0.0001$
waist	1.69997	0.0743	22.9	$\leq 0.0001$

The regression table gives the coefficients (once you find them in the middle of all this other information), so we can see that the regression equation is

$$\widehat{\%BF} = -42.73 + 1.7 \text{ Waist}$$

and that the  $R^2$  for the regression is 67.8%. (Is accounting for 68% of the variation in %Body Fat good enough to be useful? Is a prediction ME of more than 9% good enough? Health professionals might not be satisfied.)

The column of  $t$ -ratios gives the test statistics for the respective null hypotheses that the true values of the coefficients are zero. The corresponding  $P$ -values are also usually reported.

## EXERCISES

- T** 1. **Hurricane predictions.** In Chapter 7 we looked at data from the National Oceanic and Atmospheric Administration about their success in predicting hurricane tracks.

Here is a scatterplot of the error (in nautical miles) for predicting hurricane locations 72 hours in the future vs. the year in which the prediction (and the hurricane) occurred:

## REGRESSION ANALYSIS ON THE COMPUTER

All statistics packages make a table of results for a regression. These tables differ slightly from one package to another, but all are essentially the same. We've seen two examples of such tables already.

All packages offer analyses of the residuals. With some, you must request plots of the residuals as you request the regression. Others let you find the regression first and then analyze the residuals afterward. Either way, your analysis is not complete if you don't check the residuals with a histogram or Normal probability plot and a scatterplot of the residuals against  $x$  or the predicted values.

You should, of course, always look at the scatterplot of your two variables before computing a regression.

Regressions are almost always found with a computer or calculator. The calculations are too long to do conveniently by hand for data sets of any reasonable size. No matter how the regression is computed, the results are usually presented in a table that has a standard form. Here's a portion of a typical regression results table, along with annotations showing where the numbers come from:

**AS** **Activity: Regression on the Computer.** How fast is the universe expanding? And how old is it? A prominent astronomer used regression to astound the scientific community. Read the story, analyze the data, and interactively learn about each of the numbers in a typical computer regression output table.

Variable	Coefficient	SE(Coeff)	t-ratio	Prob
Constant	-42.7341	2.717	-15.7	$\leq 0.0001$
waist	1.69997	0.0743	22.9	$\leq 0.0001$

The regression table gives the coefficients (once you find them in the middle of all this other information), so we can see that the regression equation is

$$\widehat{\%BF} = -42.73 + 1.7 \text{ Waist}$$

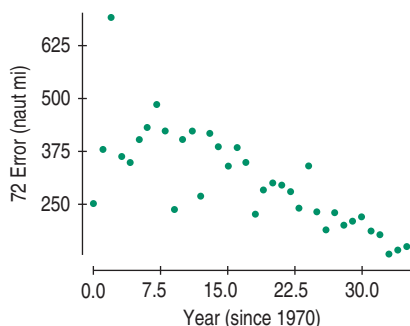
and that the  $R^2$  for the regression is 67.8%. (Is accounting for 68% of the variation in %Body Fat good enough to be useful? Is a prediction ME of more than 9% good enough? Health professionals might not be satisfied.)

The column of  $t$ -ratios gives the test statistics for the respective null hypotheses that the true values of the coefficients are zero. The corresponding  $P$ -values are also usually reported.

## EXERCISES

- T** 1. **Hurricane predictions.** In Chapter 7 we looked at data from the National Oceanic and Atmospheric Administration about their success in predicting hurricane tracks.

Here is a scatterplot of the error (in nautical miles) for predicting hurricane locations 72 hours in the future vs. the year in which the prediction (and the hurricane) occurred:



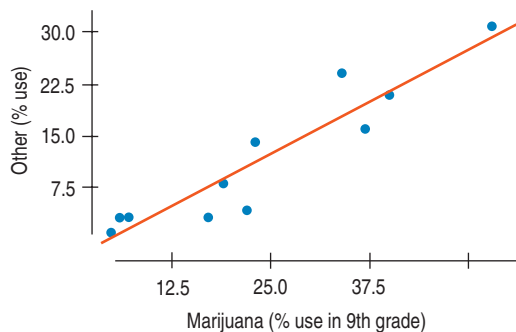
In Chapter 7 we could describe this relationship only in general terms. Now we can learn more. Here is the regression analysis:

Dependent variable is: 72Error  
 R squared = 58.5%  
 $s = 75.38$  with 36 - 2 = 34 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	453.223	24.61	18.4	$\leq 0.0001$
Year since 1970	-8.37084	1.209	-6.92	$\leq 0.0001$

- Explain in context what the regression says.
- State the hypothesis about the slope (both numerically and in words) that describes how hurricane prediction quality has changed.
- Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion in context.
- Explain what R-squared means in context.

- T** 2. **Drug use.** The *European School Study Project on Alcohol and Other Drugs*, published in 1995, investigated the use of marijuana and other drugs. Data from 11 countries are summarized in the following scatterplot and regression analysis. They show the association between the percentage of a country's ninth graders who report having smoked marijuana and who have used other drugs such as LSD, amphetamines, and cocaine.

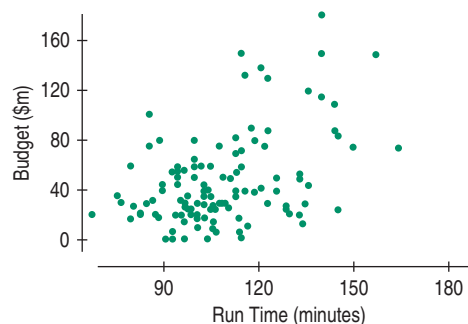


Dependent variable is: Other  
 R-squared = 87.3%  
 $s = 3.853$  with 11 - 2 = 9 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-3.06780	2.204	-1.39	0.1974
Marijuana	0.615003	0.0784	7.85	$< 0.0001$

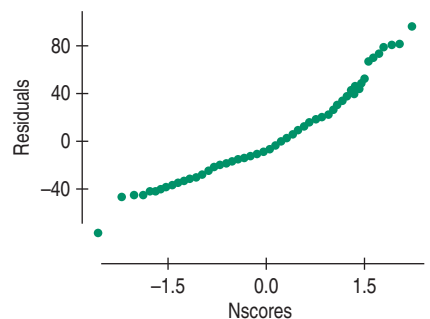
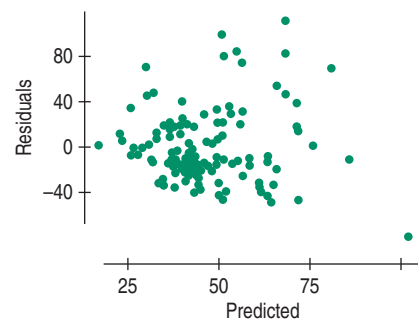
- Explain in context what the regression says.
- State the hypothesis about the slope (both numerically and in words) that describes how use of marijuana is associated with other drugs.
- Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion in context.
- Explain what R-squared means in context.
- Do these results indicate that marijuana use leads to the use of harder drugs? Explain.

- T** 3. **Movie budgets.** How does the cost of a movie depend on its length? Data on the cost (millions of dollars) and the running time (minutes) for major release films of 2005 are summarized in these plots and computer output:



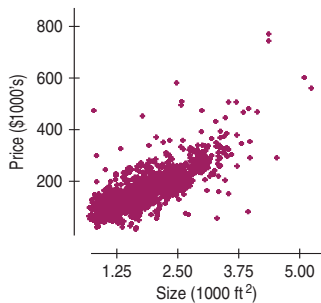
Dependent variable is: Budget(\$M)  
 R squared = 15.4%  
 $s = 32.95$  with 120 - 2 = 118 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-31.3869	17.12	-1.83	0.0693
Run Time	0.714400	0.1541	4.64	$\leq 0.0001$



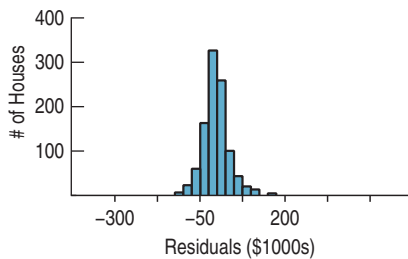
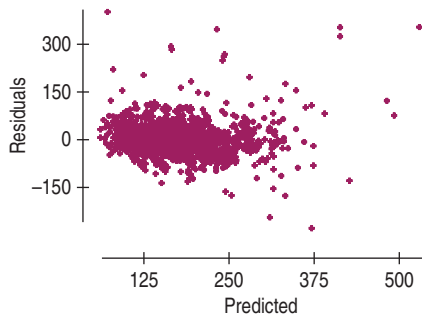
- a) Explain in context what the regression says.
- b) The intercept is negative. Discuss its value, taking note of the P-value.
- c) The output reports  $s = 32.95$ . Explain what that means in this context.
- d) What's the value of the standard error of the slope of the regression line?
- e) Explain what that means in this context.

**T 4. House prices.** How does the price of a house depend on its size? Data from Saratoga, New York, on 1064 randomly selected houses that had been sold include data on price (\$1000's) and size (1000's ft<sup>2</sup>), producing the following graphs and computer output:



Dependent variable is: Price  
 R squared = 59.5%  
 $s = 53.79$  with 1064 - 2 = 1062 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-3.11686	4.688	-0.665	0.5063
Size	94.4539	2.393	39.5	≤0.0001



- a) Explain in context what the regression says.
- b) The intercept is negative. Discuss its value, taking note of its P-value.
- c) The output reports  $s = 53.79$ . Explain what that means in this context.
- d) What's the value of the standard error of the slope of the regression line?
- e) Explain what that means in this context.

**T 5. Movie budgets: the sequel.** Exercise 3 shows computer output examining the association between the length of a movie and its cost.

- a) Check the assumptions and conditions for inference.
- b) Find a 95% confidence interval for the slope and interpret it in context.

**T 6. Second home.** Exercise 4 shows computer output examining the association between the sizes of houses and their sale prices.

- a) Check the assumptions and conditions for inference.
- b) Find a 95% confidence interval for the slope and interpret it in context.

**T 7. Hot dogs.** Healthy eating probably doesn't include hot dogs, but if you are going to have one, you'd probably hope it's low in both calories and sodium. In its July 2007 issue, *Consumer Reports* listed the number of calories and sodium content (in milligrams) for 13 brands of all-beef hot dogs it tested. Examine the association, assuming that the data satisfy the conditions for inference.

Dependent variable is: Sodium  
 R squared = 60.5%  
 $s = 59.66$  with 13 - 2 = 11 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Constant	90.9783	77.69	1.17	0.2663
Calories	2.29959	0.5607	4.10	0.0018

- a) State the appropriate hypotheses about the slope.
- b) Test your hypotheses and state your conclusion in the proper context.

**T 8. Cholesterol 2007.** Does a person's cholesterol level tend to change with age? Data collected from 1406 adults aged 45 to 62 produced the regression analysis shown. Assuming that the data satisfy the conditions for inference, examine the association between age and cholesterol level.

Dependent variable is: Chol  
 $s = 46.16$

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	194.232	13.55	14.3	≤0.0001
Age	0.771639	0.2574	3.00	0.0056

- a) State the appropriate hypothesis for the slope.
- b) Test your hypothesis and state your conclusion in the proper context.

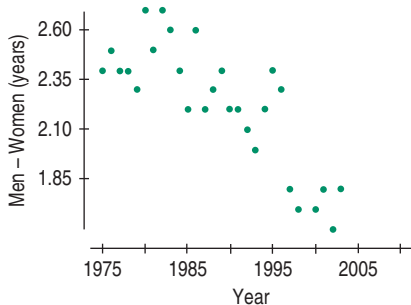
**T 9. Second frank.** Look again at Exercise 7's regression output for the calorie and sodium content of hot dogs.

- a) The output reports  $s = 59.66$ . Explain what that means in this context.
- b) What's the value of the standard error of the slope of the regression line?
- c) Explain what that means in this context.

**T 10. More cholesterol.** Look again at Exercise 8's regression output for age and cholesterol level.

- a) The output reports  $s = 46.16$ . Explain what that means in this context.
- b) What's the value of the standard error of the slope of the regression line?
- c) Explain what that means in this context.

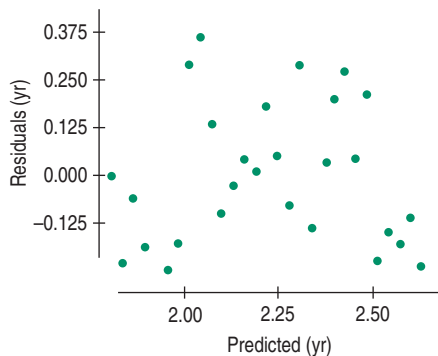
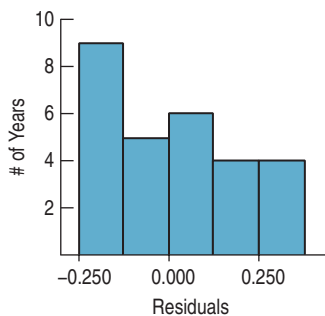
- T 11. Last dog.** Based on the regression output seen in Exercise 7, create a 95% confidence interval for the slope of the regression line and interpret your interval in context.
- T 12. Cholesterol, finis.** Based on the regression output seen in Exercise 8, create a 95% confidence interval for the slope of the regression line and interpret it in context.
- T 13. Marriage age 2003.** The scatterplot suggests a decrease in the difference in ages at first marriage for men and women since 1975. We want to examine the regression to see if this decrease is significant.



Dependent variable is: Men - Women  
 R squared = 65.6%  
 s = 0.1869 with 28 - 2 = 26 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	61.8067	8.468	7.30	≤0.0001
Year	-0.02996	0.0043	-7.04	≤0.0001

- a) Write appropriate hypotheses.
- b) Here are the residuals plot and a histogram of the residuals. Do you think the conditions for inference are satisfied? Explain.



- c) Test the hypothesis and state your conclusion about the trend in age at first marriage.

- T 14. Used cars 2007.** Classified ads in a newspaper offered several used Toyota Corollas for sale. Listed below are the ages of the cars and the advertised prices.

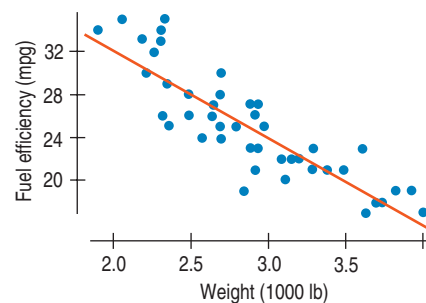
Age (yr)	Advertised Price (\$)	Age (yr)	Advertised Price (\$)
1	13990	7	6950
1	13495	7	7850
3	12999	8	6999
4	9500	8	5995
4	10495	10	4950
5	8995	10	4495
5	9495	13	2850
6	6999		

- a) Make a scatterplot for these data.
- b) Do you think a linear model is appropriate? Explain.
- c) Find the equation of the regression line.
- d) Check the residuals to see if the conditions for inference are met.

- T 15. Marriage age 2003, again.** Based on the analysis of marriage ages since 1975 given in Exercise 13, give a 95% confidence interval for the rate at which the age gap is closing. Explain what your confidence interval means.

- T 16. Used cars 2007, again.** Based on the analysis of used car prices you did for Exercise 14, create a 95% confidence interval for the slope of the regression line and explain what your interval means in context.

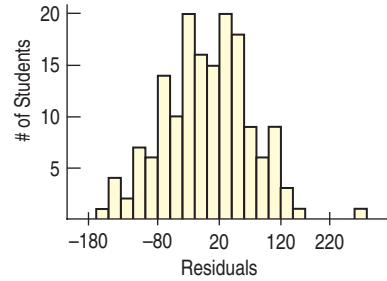
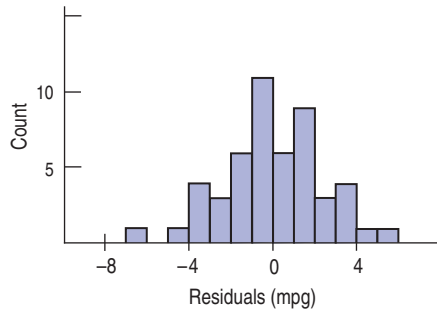
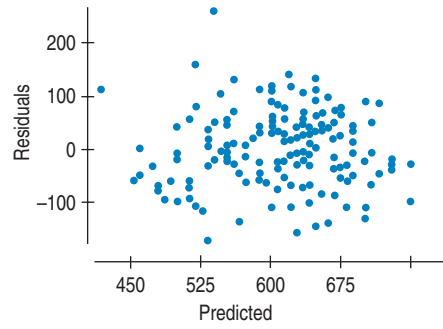
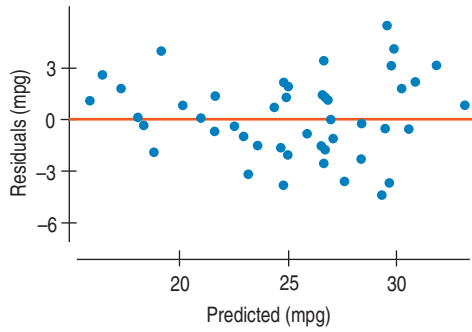
- T 17. Fuel economy.** A consumer organization has reported test data for 50 car models. We will examine the association between the weight of the car (in thousands of pounds) and the fuel efficiency (in miles per gallon). Here are the scatterplot, summary statistics, and regression analysis:



Variable	Count	Mean	StdDev
MPG	50	25.0200	4.83394
wt./1000	50	2.88780	0.511656

Dependent variable is: MPG  
 R-squared = 75.6%  
 s = 2.413 with 50 - 2 = 48 df

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	48.7393	1.976	24.7	≤0.0001
Weight	-8.21362	0.6738	-12.2	≤0.0001

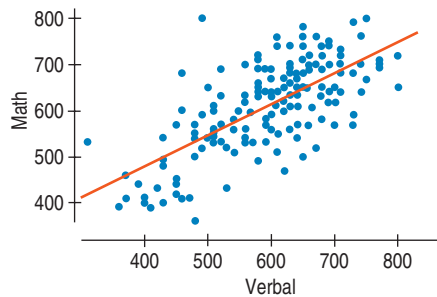


- T 18. SAT scores.** How strong was the association between student scores on the Math and Verbal sections of the old SAT? Scores on each ranged from 200 to 800 and were widely used by college admissions offices. Here are summaries and plots of the scores for a graduating class at Ithaca High School:

Variable	Count	Mean	Median	StdDev	Range	IntQRRange
Verbal	162	596.296	610	99.5199	490	140
Math	162	612.099	630	98.1343	440	150

Dependent variable is: Math  
 R-squared = 46.9%  
 $s = 71.75$  with 162  $- 2 = 160$  df

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	209.554	34.35	6.10	$\leq 0.0001$
Verbal	0.675075	0.0568	11.9	$\leq 0.0001$



- Is there evidence of an association between Math and Verbal scores? Write an appropriate hypothesis.
- Discuss the assumptions for inference.
- Test your hypothesis and state an appropriate conclusion.

- T 19. Fuel economy, part II.** Consider again the data in Exercise 17 about the gas mileage and weights of cars.
- Create a 95% confidence interval for the slope of the regression line.
  - Explain in this context what your confidence interval means.
- T 20. SATs, part II.** Consider the high school SAT scores data from Exercise 18.
- Find a 90% confidence interval for the slope of the true line describing the association between Math and Verbal scores.
  - Explain in this context what your confidence interval means.
- T 21. \*Fuel economy, part III.** Consider again the data in Exercise 17 about the gas mileage and weights of cars.
- Create a 95% confidence interval for the average fuel efficiency among cars weighing 2500 pounds, and explain what your interval means.
  - Create a 95% prediction interval for the gas mileage you might get driving your new 3450-pound SUV, and explain what that interval means.
- T 22. \*SATs again.** Consider the high school SAT scores data from Exercise 18 once more.
- Find a 90% confidence interval for the mean SAT-Math score for all students with an SAT-Verbal score of 500.
  - Find a 90% prediction interval for the Math score of the senior class president if you know she scored 710 on the Verbal section.



**T 23. Cereal.** A healthy cereal should be low in both calories and sodium. Data for 77 cereals were examined and judged acceptable for inference. The 77 cereals had between 50 and 160 calories per serving and between 0 and 320 mg of sodium per serving. Here's the regression analysis:

Dependent variable is: Sodium

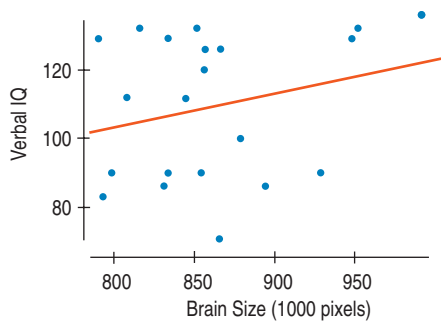
R-squared = 9.0%

s = 80.49 with 77 - 2 = 75 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	21.4143	51.47	0.416	0.6786
Calories	1.29357	0.4738	2.73	0.0079

- Is there an association between the number of calories and the sodium content of cereals? Explain.
- Do you think this association is strong enough to be useful? Explain.

**T 24. Brain size.** Does your IQ depend on the size of your brain? A group of female college students took a test that measured their verbal IQs and also underwent an MRI scan to measure the size of their brains (in 1000s of pixels). The scatterplot and regression analysis are shown, and the assumptions for inference were satisfied.



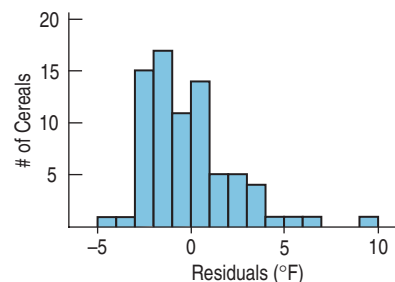
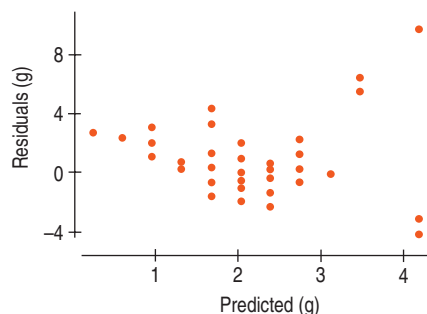
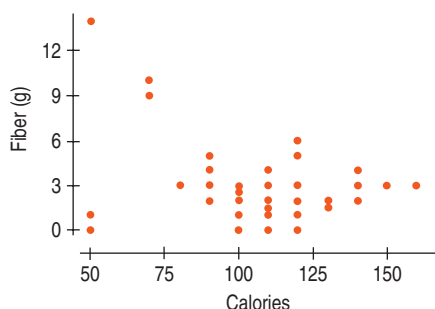
Dependent variable is: IQ\_V erbal

R-squared = 6.5%

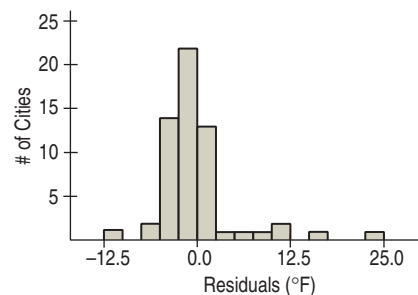
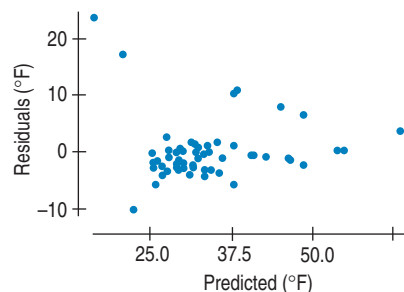
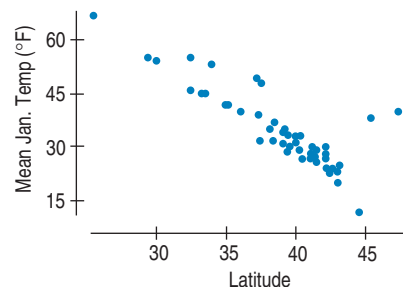
Variable	Coefficient	SE(Coeff)
Intercept	24.1835	76.38
Size	0.098842	0.0884

- Test an appropriate hypothesis about the association between brain size and IQ.
- State your conclusion about the strength of this association.

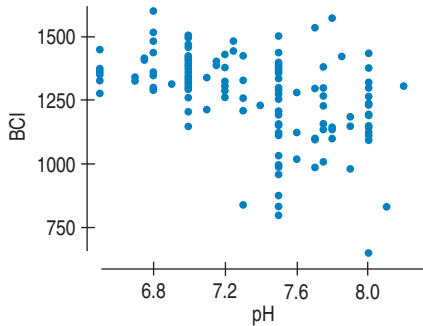
**T 25. Another bowl.** Further analysis of the data for the breakfast cereals in Exercise 23 looked for an association between *Fiber* content and *Calories* by attempting to construct a linear model. Here are several graphs. Which of the assumptions for inference are violated? Explain.



**T 26. Winter.** The output shows an attempt to model the association between average *January Temperature* (in degrees Fahrenheit) and *Latitude* (in degrees north of the equator) for 59 U.S. cities. Which of the assumptions for inference do you think are violated? Explain.



- T 27. Acid rain.** Biologists studying the effects of acid rain on wildlife collected data from 163 streams in the Adirondack Mountains. They recorded the *pH* (acidity) of the water and the *BCI*, a measure of biological diversity. Here's a scatterplot of *BCI* against *pH*:



And here is part of the regression analysis:

Dependent variable is: BCI  
 R-squared = 27.1%  
 $s = 140.4$  with 163 - 2 = 161 degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	2733.37	187.9
pH	-197.694	25.57

- State the null and alternative hypotheses under investigation.
  - Assuming that the assumptions for regression inference are reasonable, find the *t*- and *P*-values.
  - State your conclusion.
- T 28. El Niño.** Concern over the weather associated with El Niño has increased interest in the possibility that the climate on earth is getting warmer. The most common theory relates an increase in atmospheric levels of carbon dioxide ( $\text{CO}_2$ ), a greenhouse gas, to increases in temperature. Here is part of a regression analysis of the mean annual  $\text{CO}_2$  concentration in the atmosphere, measured in parts per million (ppm), at the top of Mauna Loa in Hawaii and the mean annual air temperature over both land and sea across the globe, in degrees Celsius. The scatterplots and residuals plots indicated that the data were appropriate for inference.
- Dependent variable is: *T emp*  
 R-squared = 33.4%  
 $s = 0.0809$  with 37 - 2 = 35 degrees of freedom
- | Variable      | Coefficient | SE(Coeff) |
|---------------|-------------|-----------|
| Intercept     | 15.3066     | 0.3139    |
| $\text{CO}_2$ | 0.004       | 0.0009    |
- Write the equation of the regression line.
  - Is there evidence of an association between  $\text{CO}_2$  level and global temperature?
  - Do you think predictions made by this regression will be very accurate? Explain.
- 29. Ozone.** The Environmental Protection Agency is examining the relationship between the ozone level (in parts per million) and the population (in millions) of U.S. cities. Part of the regression analysis is shown.

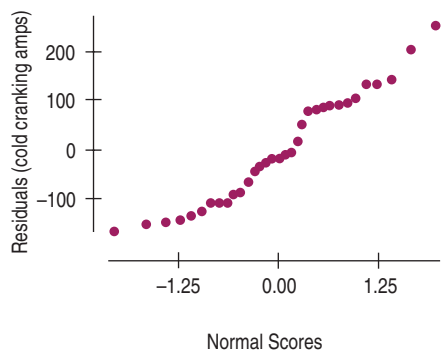
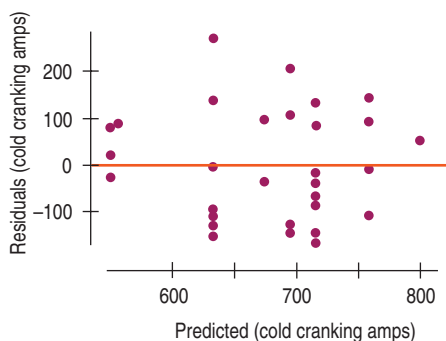
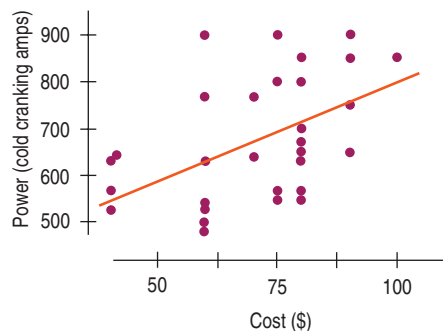
Dependent variable is: Ozone  
 R-squared = 84.4%  
 $s = 5.454$  with 16 - 2 = 14 df

Variable	Coefficient	SE(Coeff)
Intercept	18.892	2.395
Pop	6.650	1.910

- We suspect that the greater the population of a city, the higher its ozone level. Is the relationship significant? Assuming the conditions for inference are satisfied, test an appropriate hypothesis and state your conclusion in context.
  - Do you think that the population of a city is a useful predictor of ozone level? Use the values of both  $R^2$  and  $s$  in your explanation.
- T 30. Sales and profits.** A business analyst was interested in the relationship between a company's sales and its profits. She collected data (in millions of dollars) from a random sample of Fortune 500 companies and created the regression analysis and summary statistics shown. The assumptions for regression inference appeared to be satisfied.
- |          | Profits | Sales      | Dependent variable is: Profits |                    |
|----------|---------|------------|--------------------------------|--------------------|
| Count    | 79      | 79         | R-squared = 66.2%              | $s = 466.2$        |
| Mean     | 209.839 | 4178.29    | <b>Variable</b>                | <b>Coefficient</b> |
| Variance | 635,172 | 49,163,000 | Intercept                      | -176.644           |
| Std Dev  | 796.977 | 7011.63    | Sales                          | 0.092498           |
|          |         |            |                                | 0.0075             |
- Is there a significant association between sales and profits? Test an appropriate hypothesis and state your conclusion in context.
  - Do you think that a company's sales serve as a useful predictor of its profits? Use the values of both  $R^2$  and  $s$  in your explanation.
- 31. Ozone, again.** Consider again the relationship between the population and ozone level of U.S. cities that you analyzed in Exercise 29.
- Give a 90% confidence interval for the approximate increase in ozone level associated with each additional million city inhabitants.
  - \*b) For the cities studied, the mean population was 1.7 million people. The population of Boston is approximately 0.6 million people. Predict the mean ozone level for cities of that size with an interval in which you have 90% confidence.
- T 32. More sales and profits.** Consider again the relationship between the sales and profits of Fortune 500 companies that you analyzed in Exercise 30.
- Find a 95% confidence interval for the slope of the regression line. Interpret your interval in context.
  - \*b) Last year the drug manufacturer Eli Lilly, Inc., reported gross sales of \$9 billion (that's \$9,000 million). Create a 95% prediction interval for the company's profits, and interpret your interval in context.
- T 33. Start the car!** In October 2002, *Consumer Reports* listed the price (in dollars) and power (in cold cranking amps) of auto batteries. We want to know if more expensive batteries are generally better in terms of starting power. Here are several software displays:

Dependent variable is: Power  
 R-squared = 25.2%  
 s = 116.0 with 33 - 2 = 31 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	384.594	93.55	4.11	0.0003
Cost	4.14649	1.282	3.23	0.0029



- How many batteries were tested?
- Are the conditions for inference satisfied? Explain.
- Is there evidence of an association between the cost and cranking power of auto batteries? Test an appropriate hypothesis and state your conclusion.
- Is the association strong? Explain.
- What is the equation of the regression line?
- Create a 90% confidence interval for the slope of the true line.
- Interpret your interval in this context.

**T 34. Crawling.** Researchers at the University of Denver Infant Study Center wondered whether temperature might influence the age at which babies learn to crawl. Perhaps the extra clothing that babies wear in cold weather would restrict movement and delay the age at which

they started crawling. Data were collected on 208 boys and 206 girls. Parents reported the month of the baby's birth and the age (in weeks) at which their child first crawled. The table gives the average *Temperature* (°F) when the babies were 6 months old and average *Crawling Age* (in weeks) for each month of the year. Make the plots and compute the analyses necessary to answer the following questions.

Birth Month	6-Month Temperature	Average Crawling Age
Jan.	66	29.84
Feb.	73	30.52
Mar.	72	29.70
April	63	31.84
May	52	28.58
June	39	31.44
July	33	33.64
Aug.	30	32.82
Sept.	33	33.83
Oct.	37	33.35
Nov.	48	33.38
Dec.	57	32.32

- Would this association appear to be weaker, stronger, or the same if data had been plotted for individual babies instead of using monthly averages? Explain.
- Is there evidence of an association between *Temperature* and *Crawling Age*? Test an appropriate hypothesis and state your conclusion. Don't forget to check the assumptions.
- Create and interpret a 95% confidence interval for the slope of the true relationship.

**T 35. Body fat.** Do the data shown in the table below indicate an association between *Waist size* and *%Body Fat*?

- Test an appropriate hypothesis and state your conclusion.
- \*b) Give a 95% confidence interval for the mean *%Body Fat* found in people with 40-inch *Waists*.

Waist (in.)	Weight (lb)	Body Fat (%)	Waist (in.)	Weight (lb)	Body Fat (%)
32	175	6	33	188	10
36	181	21	40	240	20
38	200	15	36	175	22
33	159	6	32	168	9
39	196	22	44	246	38
40	192	31	33	160	10
41	205	32	41	215	27
35	173	21	34	159	12
38	187	25	34	146	10
38	188	30	44	219	28

- T 36. Body fat, again.** Use the data from Exercise 35 to examine the association between *Weight* and *%Body Fat*.
- Find a 90% confidence interval for the slope of the regression line of *%Body Fat* on *Weight*.
  - Interpret your interval in context.
  - Give a 95% prediction interval for the *%Body Fat* of an individual who weighs 165 pounds.
- T 37. Grades.** The data set below shows midterm scores from an Introductory Statistics course.

First Name	Midterm 1	Midterm 2	Homework
Timothy	82	30	61
Karen	96	68	72
Verena	57	82	69
Jonathan	89	92	84
Elizabeth	88	86	84
Patrick	93	81	71
Julia	90	83	79
Thomas	83	21	51
Marshall	59	62	58
Justin	89	57	79
Alexandra	83	86	78
Christopher	95	75	77
Justin	81	66	66
Miguel	86	63	74
Brian	81	86	76
Gregory	81	87	75
Kristina	98	96	84
Timothy	50	27	20
Jason	91	83	71
Whitney	87	89	85
Alexis	90	91	68
Nicholas	95	82	68
Amandeep	91	37	54
Irena	93	81	82
Yvon	88	66	82
Sara	99	90	77
Annie	89	92	68
Benjamin	87	62	72
David	92	66	78
Josef	62	43	56
Rebecca	93	87	80
Joshua	95	93	87
Ian	93	65	66
Katharine	92	98	77
Emily	91	95	83
Brian	92	80	82
Shad	61	58	65
Michael	55	65	51
Israel	76	88	67
Iris	63	62	67

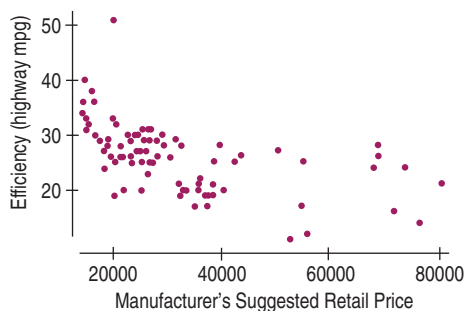
First Name	Midterm 1	Midterm 2	Homework
Mark	89	66	72
Peter	91	42	66
Catherine	90	85	78
Christina	75	62	72
Enrique	75	46	72
Sarah	91	65	77
Thomas	84	70	70
Sonya	94	92	81
Michael	93	78	72
Wesley	91	58	66
Mark	91	61	79
Adam	89	86	62
Jared	98	92	83
Michael	96	51	83
Kathryn	95	95	87
Nicole	98	89	77
Wayne	89	79	44
Elizabeth	93	89	73
John	74	64	72
Valentin	97	96	80
David	94	90	88
Marc	81	89	62
Samuel	94	85	76
Brooke	92	90	86

- Fit a model predicting the second midterm score from the first.
  - Comment on the model you found, including a discussion of the assumptions and conditions for regression. Is the coefficient for the slope statistically significant?
  - A student comments that because the P-value for the slope is very small, Midterm 2 is very well predicted from Midterm 1. So, he reasons, next term the professor can give just one midterm. What do you think?
- T 38. Grades?** The professor teaching the Introductory Statistics class discussed in Exercise 37 wonders whether performance on homework can accurately predict midterm scores.
- To investigate it, she fits a regression of the sum of the two midterms scores on homework scores. Fit the regression model.
  - Comment on the model including a discussion of the assumptions and conditions for regression. Is the coefficient for the slope “statistically significant”?
  - Do you think she can accurately judge a student’s performance without giving the midterms? Explain.
- T 39. Strike two.** Remember the Little League instructional video discussed in Chapter 25? Ads claimed it would improve the performances of Little League pitchers. To test this claim, 20 Little Leaguers threw 50 pitches each,

and we recorded the number of strikes. After the players participated in the training program, we repeated the test. The table shows the number of strikes each player threw before and after the training. A test of paired differences failed to show that this training improves ability to throw strikes. Is there any evidence that the effectiveness of the video (*After* – *Before*) depends on the player’s initial ability to throw strikes (*Before*)? Test an appropriate hypothesis and state your conclusion. Propose an explanation for what you find.

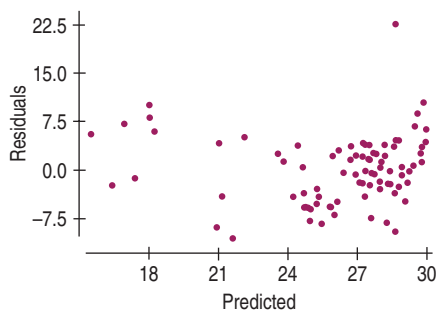
Number of Strikes (out of 50)			
Before	After	Before	After
28	35	33	33
29	36	33	35
30	32	34	32
32	28	34	30
32	30	34	33
32	31	35	34
32	32	36	37
32	34	36	33
32	35	37	35
33	36	37	32

**T 40. All the efficiency money can buy.** A sample of 84 model-2004 cars from an online information service was examined to see how fuel efficiency (as highway mpg) relates to the cost (Manufacturer’s Suggested Retail Price in dollars) of cars. Here are displays and computer output:



Dependent variable is: Highway MPG  
 R squared = 30.1%  
 s = 5.298 with 84 – 2 = 82 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Constant	33.0581	1.299	25.5	≤0.0001
MSRP	-2.16543e-4	0.0000	-5.95	≤0.0001



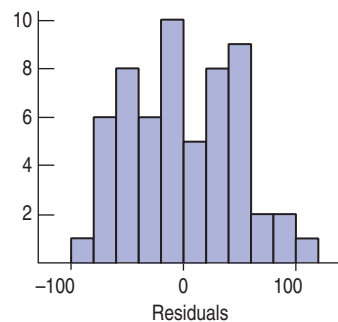
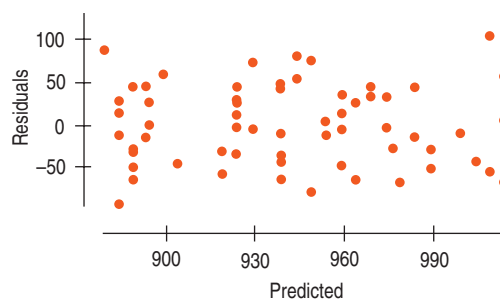
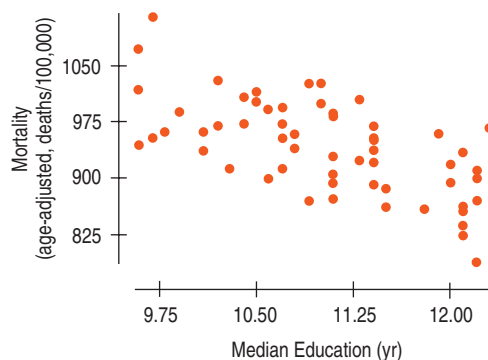
- State what you want to know, identify the variables, and give the appropriate hypotheses.
- Check the assumptions and conditions.
- If the conditions are met, complete the analysis.

**T 41. Education and mortality.** The software output below is based on the mortality rate (deaths per 100,000 people) and the education level (average number of years in school) for 58 U.S. cities.

Variable	Count	Mean	StdDev
Mortality	58	942.501	61.8490
Education	58	11.0328	0.793480

Dependent variable is: Mortality  
 R-squared = 41.0%  
 s = 47.92 with 58 – 2 = 56 degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	1493.26	88.48
Education	-49.9202	8.000



- Comment on the assumptions for inference.
- Is there evidence of a strong association between the level of *Education* in a city and the *Mortality* rate? Test an appropriate hypothesis and state your conclusion.

- c) Can we conclude that getting more education is likely (on average) to prolong your life? Why or why not?
- d) Find a 95% confidence interval for the slope of the true relationship.
- e) Explain what your interval means.
- \*f) Find a 95% confidence interval for the average *Mortality* rate in cities where the adult population completed an average of 12 years of school.

**T 42. Property assessments.** The software outputs below provide information about the *Size* (in square feet) of 18 homes in Ithaca, New York, and the city's assessed *Value* of those homes.

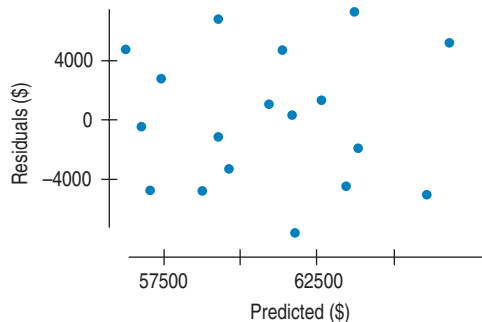
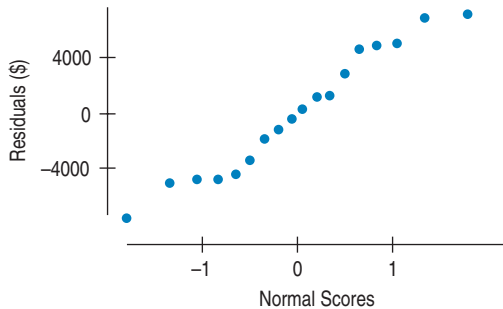
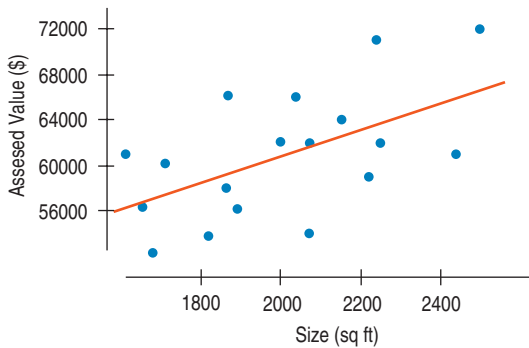
Variable	Count	Mean	StdDev	Range
Size	18	2003.39	264.727	890
Value	18	60946.7	5527.62	19710

Dependent variable is: V alue

R-squared = 32.5%

s = 4682 with 18 - 2 = 16 degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	37108.8	8664
Size	11.8987	4.290



- a) Explain why inference for linear regression is appropriate with these data.
- b) Is there a significant association between the *Size* of a home and its assessed *Value*? Test an appropriate hypothesis and state your conclusion.
- c) What percentage of the variability in assessed *Value* is explained by this regression?
- d) Give a 90% confidence interval for the slope of the true regression line, and explain its meaning in the proper context.
- e) From this analysis, can we conclude that adding a room to your house will increase its assessed *Value*? Why or why not?
- \*f) The owner of a home measuring 2100 square feet files an appeal, claiming that the \$70,200 assessed *Value* is too high. Do you agree? Explain your reasoning.



### JUST CHECKING Answers

1. A high *t*-ratio of 3.27 indicates that the slope is different from zero—that is, that there is a linear relationship between height and mouth size. The small *P*-value says that a slope this large would be very unlikely to occur by chance if, in fact, there was no linear relationship between the variables.
2. Not really. The  $R^2$  for this regression is only 15.3%, so height doesn't account for very much of the variability in mouth size.
3. The value of *s* tells the standard deviation of the residuals. Mouth sizes have a mean of 60.3 cubic centimeters. A standard deviation of 15.7 in the residuals indicates that the errors made by this regression model can be quite large relative to what we are estimating. Errors of 15 to 30 cubic centimeters would be common.

## REVIEW OF PART VII

## Inference When Variables Are Related

## Quick Review

With these last two chapters, you have added important analytical tools to your ways of looking at data. Here's a brief summary of those key concepts and skills, as well as an overview of statistical inference:

- ▶ Inferences about distributions of counts use chi-square models.
  - To see if an observed distribution is consistent with a proposed model, use a goodness-of-fit test.
  - To see if two or more observed distributions could have arisen from populations with the same model, use a test of homogeneity.
- ▶ Inference about association between two variables tests the hypothesis that it is plausible to consider the variables independent.
  - If the variables are categorical, display the data in a contingency table and use a chi-square test of independence.
  - If the variables are quantitative, display them with a scatterplot. You may use a linear regression  $t$ -test if there appears to be a linear association for which the residuals are random, consistent in terms of spread, and approximately Normal.
- ▶ You can now use statistical inference to answer questions about means, proportions, distributions, and associations.
  - No inference procedure is valid unless the underlying assumptions are true. Always check the conditions before proceeding. Many of those checks should be made by examining a graph.
  - You can make inferences about a single proportion or the difference of two proportions using Normal models.
    - You can make inferences about one mean, the difference of two independent means, or the mean of paired differences using  $t$ -models.
    - You can make inferences about distributions using chi-square models.
    - You can make inferences about associations between categorical variables using chi-square models.
    - You can make inferences about linear associations between quantitative variables using  $t$ -models.

If you look back at where we've been in this book, you'll see that statistical inference relies on almost everything we've seen. In Chapters 12 and 13 we learned techniques of collecting data using randomization—that's what makes inference possible at all. In Chapters 3, 4, and 7 we learned to plot our data and to look for the patterns and relationships we use to check the conditions that allow inference. In Chapters 3, 5, and 8 we learned about the summary statistics we use to do the mechanics of inference. We use our knowledge of randomness and probability from Chapters 11, 14, and 15 to help us think clearly about uncertainty, and the probability models of Chapters 6, 16, and 17 to measure our uncertainty precisely. Ultimately, the Central Limit Theorem of Chapter 18 makes all of inference possible.

Remember (have we said this often enough yet?): Never use any inference procedure without first checking the assumptions and conditions. On the next page we summarize the new types of inference procedures, the corresponding formulas, and the assumptions and conditions. You'll find complete summaries of all our inference procedures inside the back cover of the book. Have a look. Then you'll be ready for more opportunities to practice using these concepts and skills. . . .

### Quick Guide to Inference

Think			Show			Tell?	
Inference about?	One group or two?	Procedure	Model	Parameter	Estimate	SE	Chapter
Distributions (one categorical variable)	One sample	Goodness-of-Fit	$\chi^2$ $df = \text{cells} - 1$	$\sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$			26
	Many independent groups	Homogeneity $\chi^2$ Test					
Independence (two categorical variables)	One sample	Independence $\chi^2$ Test	$\chi^2$ $df = (r - 1)(c - 1)$				
Association (two quantitative variables)	One sample	Linear Regression $t$ -Test or Confidence Interval for $\beta$	$t$ $df = n - 2$	$\beta_1$	$b_1$	$\frac{s_e}{s_x \sqrt{n - 1}}$ (compute with technology)	27
		Confidence Interval for $\mu_v$		$\mu_v$	$y_v$	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n}}$	
		Prediction Interval for $y_v$		$y_v$	$y_v$	$\sqrt{SE^2(b_1) \cdot (x_v - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	

#### Assumptions for Inference

#### And the Conditions That Support or Override Them

##### Distributions/Association ( $\chi^2$ )

- **Goodness-of-fit** ( $df = \# \text{ of cells} - 1$ ; one variable, one sample compared with population model)
  1. Data are counts.
  2. Data in sample are independent.
  3. Sample is sufficiently large.
- **Homogeneity** [ $df = (r - 1)(c - 1)$ ; samples from many populations compared on one variable]
  1. Data are counts.
  2. Data in groups are independent.
  3. Groups are sufficiently large.
- **Independence** [ $df = (r - 1)(c - 1)$ ; sample from one population classified on two variables]
  1. Data are counts.
  2. Data are independent.
  3. Sample is sufficiently large.

##### Regression ( $t$ , $df = n - 2$ )

- **Association** between two quantitative variables ( $\beta = 0$ ?)
    1. Form of relationship is linear.
    2. Errors are independent.
    3. Variability of errors is constant.
    4. Errors have a Normal model.
1. Scatterplot looks approximately linear.
  2. No apparent pattern in residuals plot.
  3. Residuals plot has consistent spread.
  4. Histogram of residuals is approximately unimodal and symmetric or Normal probability plot reasonably straight.\*

(\*less critical as  $n$  increases)



## REVIEW EXERCISES

1. **Genetics.** Two human traits controlled by a single gene are the ability to roll one's tongue and whether one's ear lobes are free or attached to the neck. Genetic theory says that people will have neither, one, or both of these traits in the ratio 1:3:3:9 (1 attached, noncurling; 3 attached, curling; 3 free, noncurling; 9 free, curling). An Introductory Biology class of 122 students collected the data shown. Are they consistent with the genetic theory? Test an appropriate hypothesis and state your conclusion.

	Trait			
	Attached, noncurling	Attached, curling	Free, noncurling	Free, curling
Count	10	22	31	59

- T 2. **Tableware.** Nambe Mills manufactures plates, bowls, and other tableware made from an alloy of several metals. Each item must go through several steps, including polishing. To better understand the production process and its impact on pricing, the company checked the polishing time (in minutes) and the retail price (in US\$) of these items. The regression analysis is shown below. The scatterplot showed a linear pattern, and residuals were deemed suitable for inference.

Dependent variable is: Price

R-squared = 84.5%

$s = 20.50$  with 59  $- 2 = 57$  degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	-2.89054	5.730
Time	2.49244	0.1416

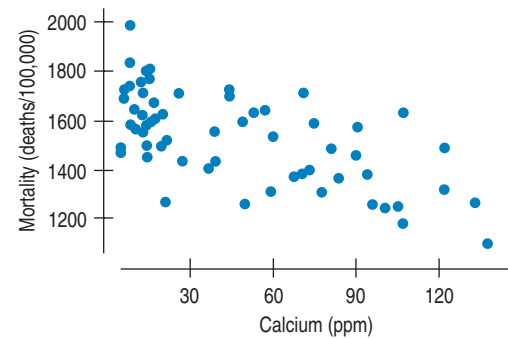
- a) How many different products were included in this analysis?
- b) What fraction of the variation in retail price is explained by the polishing time?
- c) Create a 95% confidence interval for the slope of this relationship.
- d) Interpret your interval in this context.
- T 3. **Hard water.** In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, or ppm) in the drinking water. Here are the scatterplot and regression analysis of the relationship between mortality and calcium concentration.

Dependent variable is: mortality

R-squared = 43%

$s = 143.0$  with 61  $- 2 = 59$  degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	1676	29.30
calcium	-3.23	0.48



- a) Is there an association between the hardness of the water and the mortality rate? Write the appropriate hypothesis.
- b) Assuming the assumptions for regression inference are met, what do you conclude?
- c) Create a 95% confidence interval for the slope of the true line relating calcium concentration and mortality.
- d) Interpret your interval in context.

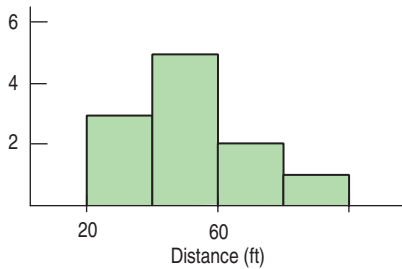
- T 4. **Mutual funds.** In March 2002, *Consumer Reports* listed the rate of return for several large-cap mutual funds over the previous 3-year and 5-year periods. ("Large cap" refers to companies worth over \$10 billion.)
- a) Create a 95% confidence interval for the difference in rate of return for the 3- and 5-year periods covered by these data. Clearly explain what your interval means.
- b) It's common for advertisements to carry the disclaimer "Past returns may not be indicative of future performance," but do these data indicate that there was an association between 3-year and 5-year rates of return?

Fund Name	Annualized Returns (%)	
	3-year	5-year
Ameristock	7.9	17.1
Clipper	14.1	18.2
Credit Suisse Strategic Value	5.5	11.5
Dodge & Cox Stock	15.2	15.7
Excelsior Value	13.1	16.4
Harbor Large Cap Value	6.3	11.5
ICAP Discretionary Equity	6.6	11.4
ICAP Equity	7.6	12.4
Neuberger Berman Focus	9.8	13.2
PBHG Large Cap Value	10.7	18.1
Pelican	7.7	12.1
Price Equity Income	6.1	10.9
USAA Cornerstone Strategy	2.5	4.9
Vanguard Equity Income	3.5	11.3
Vanguard Windsor	11.0	11.0

5. **Resume fraud.** In 2002 the Veritas Software company found out that its chief financial officer did not actually have the MBA he had listed on his resume. They fired him, and the value of the company's stock dropped 19%. Kroll, Inc., a firm that specializes in investigating such matters, said that they believe as many as 25% of background checks might reveal false information. How many such random checks would they have to do to estimate the true percentage of people who misrepresent their backgrounds to within  $\pm 5\%$  with 98% confidence?

6. **Paper airplanes.** In preparation for a regional paper airplane competition, a student tried out her latest design. The distances her plane traveled (in feet) in 11 trial flights are given here. (The world record is an astounding 193.01 feet!) The data were 62, 52, 68, 23, 34, 45, 27, 42, 83, 56, and 40 feet. Here are some summaries:

Count	11
Mean	48.3636
Median	45
StdDev	18.0846
StdErr	5.45273
IntQR	25
25th %tile	35.5000
75th %tile	60.5000

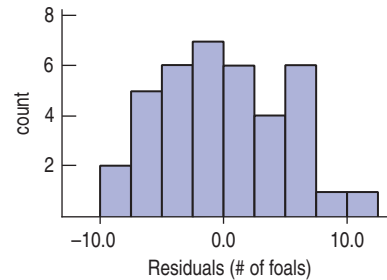
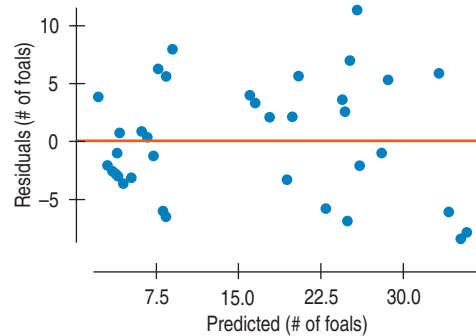
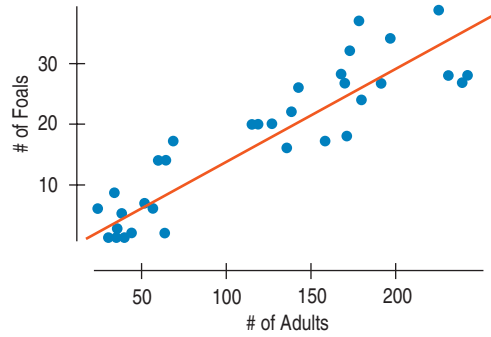


- Construct a 95% confidence interval for the true distance.
- Based on your confidence interval, is it plausible that the mean distance is 40 ft? Explain.
- How would a 99% confidence interval for the true distance differ from your answer in part a? Explain briefly, without actually calculating a new interval.
- How large a sample size would the student need to get a confidence interval half as wide as the one you got in part a, at the same confidence level?

7. **Back to Montana.** The respondents to the Montana poll described in Exercise 29 in Chapter 26 were also classified by income level: low (under \$20,000), middle (\$20,000–\$35,000), or high (over \$35,000). Is there any evidence that party enrollment there is associated with income? Test an appropriate hypothesis about this table, and state your conclusions.

	Democrat	Republican	Independent
Low	30	16	12
Middle	28	24	22
High	26	38	6

8. **Wild horses.** Large herds of wild horses can become a problem on some federal lands in the West. Researchers hoping to improve the management of these herds collected data to see if they could predict the number of foals that would be born based on the size of the current herd. Their attempt to model this herd growth is summarized in the output shown.



Variable	Count	Mean	StdDev
Adults	38	110.237	71.1809
Foals	38	15.3947	11.9945

Dependent variable is: Foals  
 R-squared = 83.5%  
 $s = 4.941$  with 38 - 2 = 36 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	-1.57835	1.492	-1.06	0.2970
Adults	0.153969	0.0114	13.5	$\leq 0.0001$

- How many herds of wild horses were studied?
- Are the conditions necessary for inference satisfied? Explain.
- Create a 95% confidence interval for the slope of this relationship.
- Explain in this context what that slope means.
- Suppose that a new herd with 80 adult horses is located. Estimate, with a 90% prediction interval, the number of foals that may be born.

9. **Lefties and music.** In an experiment to see if left- and right-handed people have different abilities in music, subjects heard a tone and were then asked to identify which of several other tones matched the first. Of 76 right-handed subjects, 38 were successful in completing this test, compared with 33 of 53 lefties. Is this strong evidence of a difference in musical abilities based on handedness?

T 10. **AP Statistics scores.** In 2001, more than 41,000 Statistics students nationwide took the Advanced Placement Examination in Statistics. The national distribution of scores and the results at Ithaca High School are shown in the table.

Score	National Distribution	Ithaca High School	
		Number of boys	Number of girls
5	11.5%	13	13
4	23.4%	21	15
3	24.9%	6	13
2	19.1%	7	3
1	21.1%	4	2

- a) Is the distribution of scores at this high school significantly different from the national results?
- b) Was there a significant difference between the performances of boys and girls at this school?

T 11. **Polling.** How accurate are pollsters in predicting the outcomes of congressional elections? The table shows the actual number of Democratic party seats in the House of Representatives and the number predicted by the Gallup organization for nonpresidential election years between World War II and 1998.

- a) Is there a significant difference between the number of seats predicted for the Democrats and the number they actually held? Test an appropriate hypothesis and state your conclusions.

Democratic Party Congressmen		
Year	Predicted	Actual
1946	190	188
1950	235	234
1954	232	232
1958	272	283
1962	259	258
1966	247	248
1970	260	255
1974	292	291
1978	277	277
1982	275	269
1986	264	258
1990	260	267
1994	201	204
1998	211	211

b) Is there a strong association between the pollsters' predictions and the outcomes of the elections? Test an appropriate hypothesis and state your conclusions.

T 12. **Twins.** In 2000 The *Journal of the American Medical Association* published a study that examined a sample of pregnancies that resulted in the birth of twins. Births were classified as preterm with intervention (induced labor or cesarean), preterm without such procedures, or term or postterm. Researchers also classified the pregnancies by the level of prenatal medical care the mother received (inadequate, adequate, or intensive). The data, from the years 1995–1997, are summarized in the table below. Figures are in thousands of births. (*JAMA* 284 [2000]: 335–341)

TWIN BIRTHS, 1995–1997 (IN THOUSANDS)					
Level of Prenatal Care		Preterm (induced or Cesarean)	Preterm (without procedures)	Term or postterm	Total
		Intensive	18	15	
Adequate	46	43	65	154	
Inadequate	12	13	38	63	
Total	76	71	131	278	

Is there evidence of an association between the duration of the pregnancy and the level of care received by the mother?

T 13. **Twins, again.** After reading of the *JAMA* study in Exercise 12, a large city hospital examined their records of twin births for several years and found the data summarized in the table below. Is there evidence that the way the hospital deals with pregnancies involving twins may have changed?



Outcome of Pregnancy	1990	1995	2000
	Preterm (induced or cesarean)	11	13
Preterm (without procedures)	13	14	18
Term or postterm	27	26	32

14. **Preemies.** Do the effects of being born prematurely linger into adulthood? Researchers examined 242 Cleveland-area children born prematurely between 1977 and 1979, and compared them with 233 children of normal birth weight; 24 of the “preemies” and 12 of the other children were described as being of “subnormal height” as adults. Is this evidence that babies born with a very low birth weight are more likely to be smaller than normal

adults? (“Outcomes in Young Adulthood for Very-Low-Birth-Weight Infants,” *New England Journal of Medicine*, 346, no. 3 [January 2002])

- T 15. LA rainfall.** The Los Angeles Almanac Web site reports recent annual rainfall (in inches), as shown in the table.
- Create a 90% confidence interval for the mean annual rainfall in LA.
  - If you wanted to estimate the mean annual rainfall with a margin of error of only 2 inches, how many years’ data would you need?
  - Do these data suggest any change in annual rainfall as time passes? Check for an association between rainfall and year.

Year	Rain (in.)	Year	Rain (in.)
1980	8.96	1991	21.00
1981	10.71	1992	27.36
1982	31.28	1993	8.14
1983	10.43	1994	24.35
1984	12.82	1995	12.46
1985	17.86	1996	12.40
1986	7.66	1997	31.01
1987	12.48	1998	9.09
1988	8.08	1999	11.57
1989	7.35	2000	17.94
1990	11.99	2001	4.42

- T 16. Age and party.** The Gallup Poll conducted a representative telephone survey during the first quarter of 1999. Among the reported results was the following table concerning the preferred political party affiliation of respondents and their ages. Is there evidence of age-based differences in party affiliation in the United States?

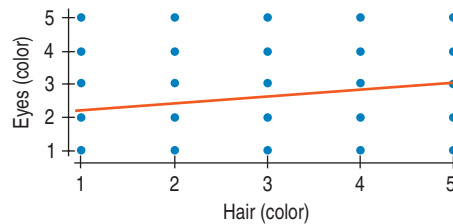
Age	Repub- lican	Democ- ratic	Inde- pendent	Total
	18–29	241	351	409
30–49	299	330	370	999
50–64	282	341	375	998
65+	279	382	343	1004
Total	1101	1404	1497	4002

- Will you conduct a test of homogeneity or independence? Why?
  - Test an appropriate hypothesis.
  - State your conclusion, including an analysis of differences you find (if any).
- T 17. Eye and hair color.** A survey of 1021 school-age children was conducted by randomly selecting children from several large urban elementary schools. Two of the questions concerned eye and hair color. In the survey, the following codes were used:

Hair Color	Eye Color
1 = Blond	1 = Blue
2 = Brown	2 = Green
3 = Black	3 = Brown
4 = Red	4 = Grey
5 = Other	5 = Other

The Statistics students analyzing the data were asked to study the relationship between eye and hair color.

- a) One group of students produced the output shown below. What kind of analysis is this? What are the null and alternative hypotheses? Is the analysis appropriate? If so, summarize the findings, being sure to include any assumptions you’ve made and/or limitations to the analysis. If it’s not an appropriate analysis, state explicitly why not.



Dependent variable is: Eyes

R-squared = 3.7%

s = 1.112 with 1021 - 2 = 1019 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	1.99541	0.08346	23.9	≤ 0.0001
Hair	0.211809	0.03372	0.28	≤ 0.0001

- b) A second group of students used the same data to produce the output shown below. The table displays counts and standardized residuals in each cell. What kind of analysis is this? What are the null and alternative hypotheses? Is the analysis appropriate? If so, summarize the findings, being sure to include any assumptions you’ve made and/or limitations to the analysis. If it’s not an appropriate analysis, state explicitly why not.

		Eye Color				
		1	2	3	4	5
Hair Color	1	143 7.67540	30 0.41799	58 -5.88169	15 -0.63925	12 -0.31451
	2	90 -2.57141	45 0.29019	215 1.72235	30 0.49189	20 -0.08246
	3	28 -5.39425	15 -2.34780	190 6.28154	10 -1.76376	10 -0.80382
	4	30 2.06116	15 2.71589	10 -4.05540	10 2.37402	5 0.75993
	5	10 -0.52195	5 0.33262	15 -0.94192	5 1.36326	5 2.07578

$$\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = 223.6 \quad \text{P-value} < 0.00001$$

- 18. Depression and the Internet.** The September 1998 issue of the *American Psychologist* published an article reporting on an experiment examining “the social and psychological impact of the Internet on 169 people in 73 households during their first 1 to 2 years online.” In the experiment, a sample of households was offered free Internet access for one or two years in return for allowing their time and activity online to be tracked. The members of the households who participated in the study were also given a battery of tests at the beginning and again at the end of the study. One of the tests measured the subjects’ levels of depression on a 4-point scale, with higher numbers meaning the person was more depressed. Internet usage was measured in average number of hours per week. The regression analysis examines the association between the subjects’ depression levels and the amounts of Internet use. The conditions for inference were satisfied.

Dependent variable is: Depression After  
 R-squared = 4.6%  
 $s = 0.4563$  with 162  $- 2 = 160$  degrees of freedom

Variable	Coefficient	SE(coeff)	t-ratio	Prob
Constant	0.565485	0.0399	14.2	$\leq 0.0001$
Intr_use	0.019948	0.0072	2.76	0.0064

- Do these data indicate that there is an association between Internet use and depression? Test an appropriate hypothesis and state your conclusion clearly.
- One conclusion of the study was that those who spent more time online tended to be more depressed at the end of the experiment. News headlines said that too much time on the Internet can lead to depression. Does the study support this conclusion? Explain.
- As noted, the subjects’ depression levels were tested at both the beginning and the end of this study; higher scores indicated the person was more depressed. Results are summarized in the table. Is there evidence that the depression level of the subjects changed during this study?

Depression Level  
 162 subjects

Variable	Mean	StdDev
DeprBfore	0.730370	0.487817
DeprAfter	0.611914	0.461932
Difference	-0.118457	0.552417

- 19. Pregnancy.** In 1998 a San Diego reproductive clinic reported 42 live births to 157 women under the age of 38, but only 7 successes for 89 clients aged 38 and older. Is this evidence of a difference in the effectiveness of the clinic’s methods for older women?
- Test the appropriate hypotheses, using the two-proportion  $z$ -procedure.
  - Repeat the analysis, using an appropriate chi-square procedure.
  - Explain how the two results are equivalent.
- 20. Eating in front of the TV.** Roper Reports asked a random sample of people in 30 countries whether they agreed with the statement “I like to nibble while reading or watching TV.” Allowable responses were “Agree completely”, “Agree somewhat”, “Neither disagree nor

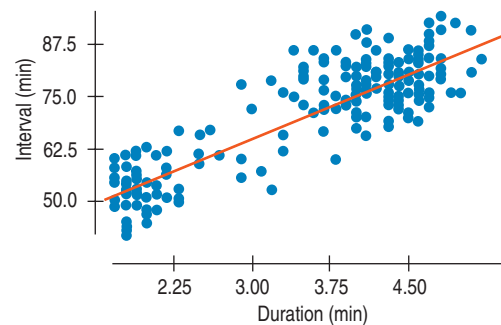
agree”, “Disagree somewhat”, “Disagree completely”, and “I Don’t Know/No Response”. Does a person’s age influence their response? Here are data from 3792 respondents in the 2006 sample of five countries (China, India, France, United Kingdom, and United States) for three age groups (Teens, 30’s (30–39) and Over 60):

	Neither				
	Agree Completely	Agree Somewhat	Disagree Nor	Disagree Somewhat	Disagree Completely
Teen	369	540	299	175	106
30’s	272	522	325	229	170
60 +	93	207	153	154	178

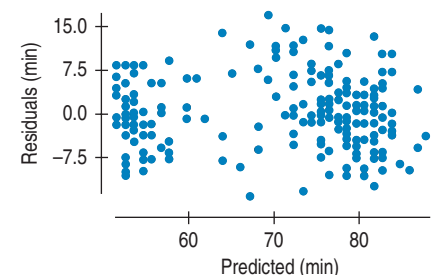
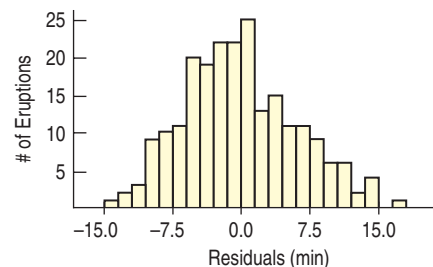
- Make an appropriate display of these data.
- Does a person’s age seem to affect their response to the question about nibbling?

- T 21. Old Faithful.** As you saw in an earlier chapter, Old Faithful isn’t all that faithful. Eruptions do not occur at uniform intervals and may vary greatly. Can we improve our chances of predicting the time of the next eruption if we know how long the previous eruption lasted?

- Describe what you see in this scatterplot.



- Write an appropriate hypothesis.
- Here are a histogram of the residuals and the residuals plot. Do you think the assumptions for inference are met? Explain.



d) State a conclusion based on this regression analysis:

Dependent variable is: Inter val

R-squared = 77.0%

s = 6.159 with 222 - 2 = 220 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	33.9668	1.428	23.8	≤ 0.0001
Duration	10.3582	0.3822	27.1	≤ 0.0001

Variable	Mean	StdDev
Duration	3.57613	1.08395
Inter val	71.0090	12.7992

- e) The second table shows the summary statistics for the two variables. Create a 95% confidence interval for the mean length of time that will elapse following a 2-minute eruption.
- f) You arrive at Old Faithful just as an eruption ends. Witnesses say it lasted 4 minutes. Create a 95% prediction interval for the length of time you will wait to see the next eruption.

22. **Togetherhness.** Are good grades in high school associated with family togetherness? A simple random sample of 142 high-school students was asked how many meals per week their families ate together. Their responses produced a mean of 3.78 meals per week, with a standard deviation of 2.2. Researchers then matched these responses against the students' grade point averages. The scatterplot appeared to be reasonably linear, so they went ahead with the regression analysis, seen below. No apparent pattern emerged in the residuals plot.

Dependent variable: GP A

R-squared = 11.0%

s = 0.6682 with 142 - 2 = 140 df

Variable	Coefficient	SE(Coeff)
Intercept	2.7288	0.1148
Meals/wk	0.1093	0.0263

- a) Is there evidence of an association? Test an appropriate hypothesis and state your conclusion.
- b) Do you think this association would be useful in predicting a student's grade point average? Explain.
- c) Are your answers to parts a and b contradictory? Explain.

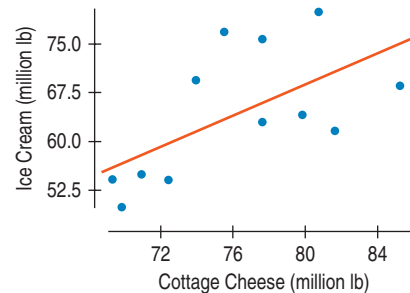
23. **Learning math.** Developers of a new math curriculum called "Accelerated Math" compared performances of students taught by their system with control groups of students in the same schools who were taught using traditional instructional methods and materials. Statistics about pretest and posttest scores are shown in the table. (J. Ysseldyke and S. Tardrew, *Differentiating Math Instruction*, Renaissance Learning, 2002)

- a) Did the groups differ in average math score at the start of this study?
- b) Did the group taught using the Accelerated Math program show a significant improvement in test scores?
- c) Did the control group show a significant improvement in test scores?
- d) Were gains significantly higher for the Accelerated Math group than for the control group?

		Instructional Method	
		Acc. math	Control
Number of students		231	245
Pretest	Mean	560.01	549.65
	St. Dev	84.29	74.68
Post-test	Mean	637.55	588.76
	St. Dev	82.9	83.24
Individual gain	Mean	77.53	39.11
	St. Dev.	78.01	66.25

24. **Pesticides.** A study published in 2002 in the journal *Environmental Health Perspectives* examined the gender ratios of children born to workers exposed to dioxin in Russian pesticide factories. The data covered the years from 1961 to 1988 in the city of Ufa, Bashkortostan, Russia. Of 227 children born to workers exposed to dioxin, only 40% were male. Overall in the city of Ufa, the proportion of males was 51.2%. Is this evidence that human exposure to dioxin may result in the birth of more girls? (An interesting note: It appeared that paternal exposure was most critical; 51% of babies born to mothers exposed to the chemical were boys.)

T 25. **Dairy sales.** Peninsula Creameries sells both cottage cheese and ice cream. The CEO recently noticed that in months when the company sells more cottage cheese, it seems to sell more ice cream as well. Two of his aides were assigned to test whether this is true or not. The first aide's plot and analysis of sales data for the past 12 months (in millions of pounds for cottage cheese and for ice cream) appear below.



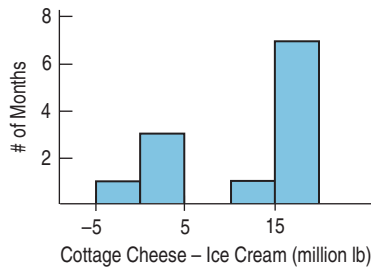
Dependent variable is: Ice cream

R-squared = 36.9%

s = 8.320 with 12 - 2 = 10 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Constant	-26.5306	37.68	-0.704	0.4975
Cottage C ...	1.19334	0.4936	2.42	0.0362

The other aide looked at the differences in sales of ice cream and cottage cheese for each month and created the following output:



Cottage Cheese	Ice Cream
Count	12
Mean	11.8000
Median	15.3500
StdDev	7.99386
IntQRRange	14.3000
25th %tile	3.20000
75th %tile	17.5000

Test H0:  $\mu[CC - IC] = 0$  vs Ha:  $\mu[CC - IC] \neq 0$   
 Sample Mean = 11.800000 t-Statistic = 5.113 w/ 11 df  
 Prob = 0.0003  
 Lower 95% bound = 6.7209429  
 Upper 95% bound = 16.879057

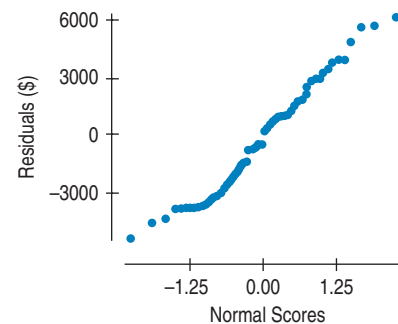
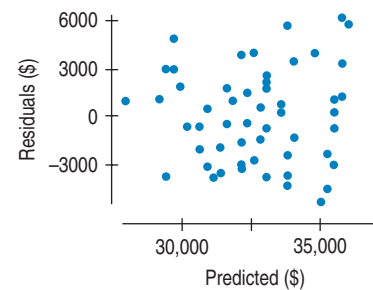
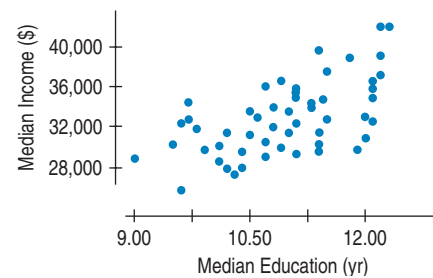
- Which analysis would you use to answer the CEO's question? Why?
- What would you tell the CEO?
- Which analysis would you use to test whether the company sells more cottage cheese or ice cream in a typical year? Why?
- What would you tell the CEO about this other result?
- What assumptions are you making in the analysis you chose in part a? What assumptions are you making in the analysis in part c?
- Next month's cottage cheese sales are 82 million pounds. Ice cream sales are not yet available. How much ice cream do you predict Peninsula Creameries will sell?
- Give a 95% confidence interval for the true slope of the regression equation of ice cream sales by cottage cheese sales.
- Explain what your interval means.

**26. Infiximab.** In an article appearing in the journal *The Lancet* in 2002, medical researchers reported on the experimental use of the arthritis drug infiximab in treating Crohn's disease. In a trial, 573 patients were given initial 5-mg injections of the drug. Two weeks later, 335 had responded positively. These patients were then randomly assigned to three groups. Group I received continued injections of a placebo, Group II continued with 5 mg of infiximab, and Group III received 10 mg of the drug. After 30 weeks, 23 of 110 Group I patients were in remission, compared with 44 of 113 Group II and 50 of 112 Group III patients. Do these data indicate that continued treatment with infiximab is of value for Crohn's disease patients who exhibit a positive initial response to the drug?

**T 27. Weight loss.** A weight loss clinic advertises that its program of diet and exercise will allow clients to lose 10 pounds in one month. A local reporter investigating weight reduction gets permission to interview a randomly selected sample of clients who report the given weight losses during their first month in this program. Create a confidence interval to test the clinic's claim that the typical weight loss is 10 pounds.

Pounds Lost	
9.5	9.5
13	9
9	8
10	7.5
11	10
9	7
5	8
9	10.5
12.5	10.5
6	9

**T 28. Education vs. income.** The information below examines the median income and education level (years in school) for several U.S. cities.



Variable	Count	Mean	StdDev
Education	57	10.9509	0.848344
Income	57	32742.6	3618.01

Dependent variable is: Income  
 R-squared = 32.9%  
 s = 2991 with 57 - 2 = 55 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	5970.05	5175	1.15	0.2537
Education	2444.79	471.2	5.19	≤ 0.0001

- Do you think the assumptions for inference are met? Explain.
- Does there appear to be an association between education and income levels in these cities?
- Would this association appear to be weaker, stronger, or the same if data were plotted for individual people rather than for cities in aggregate? Explain.
- Create and interpret a 95% confidence interval for the slope of the true line that describes the association between income and education.
- Predict the median income for cities where residents spent an average of 11 years in school. Describe your estimate with a 90% confidence interval, and interpret that result.

**T 29. Diet.** Thirteen overweight women volunteered for a study to determine whether eating specially prepared crackers before a meal could help them lose weight. The subjects were randomly assigned to eat crackers with different types of fiber (bran fiber, gum fiber, both, and a control cracker). Unfortunately, some of the women developed uncomfortable bloating and upset stomachs. Researchers suspected that some of the crackers might be at fault. The contingency table of “Cracker” versus “Bloat” shows the relationship between the four different types of crackers and the reported bloating. The study was paid for by the manufacturers of the gum fiber. What would you recommend to them about the prospects for marketing their new diet cracker?

		Bloat	
		Little/None	Moderate/Severe
Cracker	Bran	11	2
	Gum	4	9
	Combo	7	6
	Control	8	4

**T 30. Cramming.** Students in two basic Spanish classes were required to learn 50 new vocabulary words. One group of 45 students received the list on Monday and studied the words all week. Statistics summarizing this group’s scores on Friday’s quiz are given. The other group of 25 students

did not get the vocabulary list until Thursday. They also took the quiz on Friday, after “cramming” Thursday night. Then, when they returned to class the following Monday, they were retested—without advance warning. Both sets of test scores for these students are shown.

**Group 1**

**Fri.**

Number of students = 45

Mean = 43.2 (of 50)

StDev = 3.4

Students passing (score ≥ 40) = 33%

**Group 2**

	Fri.	Mon.	Fri.	Mon.
42	36	50	47	
44	44	34	34	
45	46	38	31	
48	38	43	40	
44	40	39	41	
43	38	46	32	
41	37	37	36	
35	31	40	31	
43	32	41	32	
48	37	48	39	
43	41	37	31	
45	32	36	41	
47	44			

- On Friday, did the week-long study group have a mean score significantly higher than that of the overnight crammers?
- Was there a significant difference in the percentages of students who passed the quiz on Friday?
- Is there any evidence that when students cram for a test, their “learning” does not last for 3 days?
- Use a 95% confidence interval to estimate the mean number of words that might be forgotten by crammers.
- Is there any evidence that how much students forget depends on how much they “learned” to begin with?



# Selected Formulas

$$\text{Range} = \text{Max} - \text{Min}$$

$$\text{IQR} = Q3 - Q1$$

$$\text{Outlier Rule-of-Thumb: } y < Q1 - 1.5 \times \text{IQR} \quad \text{or} \quad y > Q3 + 1.5 \times \text{IQR}$$

$$\bar{y} = \frac{\sum y}{n}$$

$$s = \sqrt{\frac{\sum (y - \bar{y})^2}{n - 1}}$$

$$z = \frac{y - \mu}{\sigma} \text{ (model based)}$$

$$z = \frac{y - \bar{y}}{s} \text{ (data based)}$$

$$r = \frac{\sum z_x z_y}{n - 1}$$

$$\hat{y} = b_0 + b_1 x \quad \text{where } b_1 = \frac{rs_y}{s_x} \text{ and } b_0 = \bar{y} - b_1 \bar{x}$$

$$P(\mathbf{A}) = 1 - P(\mathbf{A}^c)$$

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$$

$$P(\mathbf{A} \cap \mathbf{B}) = P(\mathbf{A}) \times P(\mathbf{B}|\mathbf{A})$$

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{A})}$$

If  $\mathbf{A}$  and  $\mathbf{B}$  are independent,  $P(\mathbf{B}|\mathbf{A}) = P(\mathbf{B})$

$$E(X) = \mu = \sum x \cdot P(x)$$

$$E(X \pm c) = E(X) \pm c$$

$$E(aX) = aE(X)$$

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$\text{Var}(X) = \sigma^2 = \sum (x - \mu)^2 P(x)$$

$$\text{Var}(X \pm c) = \text{Var}(X)$$

$$\text{Var}(aX) = a^2 \text{Var}(X)$$

$$\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y),$$

if  $X$  and  $Y$  are independent

$$\text{Geometric: } P(x) = q^{x-1}p \quad \mu = \frac{1}{p} \quad \sigma = \sqrt{\frac{q}{p^2}}$$

$$\text{Binomial: } P(x) = \binom{n}{x} p^x q^{n-x} \quad \mu = np \quad \sigma = \sqrt{npq}$$

$$\hat{p} = \frac{x}{n} \quad \mu(\hat{p}) = p \quad \text{SD}(\hat{p}) = \sqrt{\frac{pq}{n}}$$

Sampling distribution of  $\bar{y}$ :

(CLT) As  $n$  grows, the sampling distribution approaches the Normal model with

$$\mu(\bar{y}) = \mu_y \quad SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

**Inference:**

Confidence interval for parameter = *statistic*  $\pm$  *critical value*  $\times$  *SD(statistic)*

$$\text{Test statistic} = \frac{\text{Statistic} - \text{Parameter}}{SD(\text{statistic})}$$

Parameter	Statistic	SD(statistic)	SE(statistic)
$p$	$\hat{p}$	$\sqrt{\frac{pq}{n}}$	$\sqrt{\frac{\hat{p}\hat{q}}{n}}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$	$\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
$\mu$	$\bar{y}$	$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
$\mu_1 - \mu_2$	$\bar{y}_1 - \bar{y}_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$\mu_d$	$\bar{d}$	$\frac{\sigma_d}{\sqrt{n}}$	$\frac{s_d}{\sqrt{n}}$
$\sigma_\varepsilon$	$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$		
$\beta_1$	$b_1$		$\frac{s_e}{s_x\sqrt{n - 1}}$
$^*\mu_\nu$	$\hat{y}_\nu$		$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$
$^*y_\nu$	$\hat{y}_\nu$		$\sqrt{SE^2(b_1) \cdot (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Pooling: For testing difference between proportions:  $\hat{p}_{pooled} = \frac{y_1 + y_2}{n_1 + n_2}$

For testing difference between means:  $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

Substitute these pooled estimates in the respective SE formulas for both groups when assumptions and conditions are met.

$$\text{Chi-square: } \chi^2 = \sum \frac{(obs - exp)^2}{exp}$$

# Guide to Statistical Software

## Chapter 3. Displaying and Describing Categorical Data

### DATA DESK

To make a bar chart or pie chart, select the variable. In the **Plot** menu, choose **Bar Chart** or **Pie Chart**. To make a frequency table, in the **Calc** menu choose **Frequency Table**.

### COMMENTS

These commands treat the data as categorical even if they are numerals. If you select a quantitative variable by mistake, you'll see an error message warning of too many categories.

### EXCEL

First make a pivot table (Excel's name for a frequency table). From the **Data** menu, choose **Pivot Table** and **Pivot Chart Report**. When you reach the Layout window, drag your variable to the row area and drag your variable again to the data area. This tells Excel to count the occurrences of each category. Once you have an Excel pivot table, you can construct bar charts and pie charts. Click inside the Pivot Table.

Click the Pivot Table Chart Wizard button. Excel creates a bar chart. A longer path leads to a pie chart; see your Excel documentation.

### COMMENTS

Excel uses the pivot table to specify the category names and find counts within each category. If you already have that information, you can proceed directly to the Chart Wizard.

### EXCEL 2007

To make a bar chart:

- ▶ Select the variable in Excel you want to work with.
- ▶ Choose the **Column** command from the Insert tab in the Ribbon.
- ▶ Select the appropriate chart from the drop-down dialog.

To change the bar chart into a pie chart:

- ▶ Right-click the chart and select **Change Chart Type...** from the menu. The Chart type dialog opens.
- ▶ Select a pie chart type.
- ▶ Click the **OK** button. Excel changes your bar chart into a pie chart.

### JMP

JMP makes a bar chart and frequency table together. From the **Analyze** menu, choose **Distribution**. In the Distribution dialog, drag the name of the variable into the empty variable window beside the label "Y, Columns"; click **OK**. To make a pie chart, choose **Chart** from the **Graph** menu. In the Chart dialog, select the variable name from the Columns

list, click on the button labeled "Statistics," and select "N" from the drop-down menu. Click the "**Categories, X, Levels**" button to assign the same variable name to the X-axis. Under Options, click on the **second** button—labeled "**Bar Chart**"—and select "Pie" from the drop-down menu.

### MINITAB

To make a bar chart, choose **Bar Chart** from the **Graph** menu. Select "Counts of unique values" in the first menu, and select "Simple" for the type of graph. Click **OK**.

In the Chart dialog, enter the name of the variable that you wish to display in the box labeled "Categorical variables." Click **OK**.

**SPSS**

To make a bar chart, open the **Chart Builder** from the **Graphs** menu.  
Click the **Gallery** tab.  
Choose **Bar Chart** from the list of chart types.  
Drag the appropriate bar chart onto the canvas.

Drag a categorical variable onto the x-axis drop zone.  
Click **OK**.

**COMMENTS**

A similar path makes a pie chart by choosing **Pie chart** from the list of chart types.

**TI-NSPIRE**

The TI-Nspire Handheld does not display plots for categorical variables.

**TI-89**

The TI-89 won't do displays for categorical variables.

**Chapter 4. Displaying and Summarizing Quantitative Data****DATA DESK**

To make a histogram:

- ▶ Select the variable to display.
- ▶ In the **Plot** menu, choose **Histogram**.

To calculate summaries:

- ▶ In the **Calc** menu, open the **summaries** submenu. **Options** offer separate tables, a single unified table, and other formats.

**EXCEL**

Excel cannot make histograms or dotplots without a third-party add-in.

To calculate summaries:

Click on an empty cell. Type an equal sign and choose "**Average**" from the popup list of functions that appears to the left of the text-editing box. Enter the data range in the box that says "**Number 1.**" Click the **OK** button.

To compute the standard deviation of a column of data directly, use the **STDEV** from the popup list of functions in the same way.

**COMMENTS**

Excel's Data Analysis add-in does offer something called a histogram, but it just makes a crude frequency table, and the Chart Wizard cannot then create a statistically appropriate histogram. The DDXL add-in provided on our DVD adds these and other capabilities to Excel.

Excel's STDEV function should not be used for data values larger in magnitude than 100,000 or for lists of more than a few thousand values. It is programmed with an unstable formula that can generate rounding errors when these limits are exceeded.

**EXCEL 2007**

In Excel 2007 there is another way to find some of the standard summary statistics. For example, to compute the mean:

- ▶ Click on an empty cell.
- ▶ Go to the Formulas tab in the Ribbon. Click on the drop down arrow next to "AutoSum" and choose "**Average**".
- ▶ Enter the data range in the formula displayed in the empty box you selected earlier.
- ▶ Press **Enter**. This computes the mean for the values in that range.

To compute the standard deviation:

- ▶ Click on an empty cell.
- ▶ Go to the Formulas tab in the Ribbon and click the drop down arrow next to "AutoSum" and select "**More functions...**"
- ▶ In the dialog window that opens, select "**STDEV**" from the list of functions and click **OK**. A new dialog window opens. Enter a range of fields into the text fields and click **OK**.

Excel 2007 computes the standard deviation for the values in that range and places it in the specified cell of the spreadsheet.

**JMP**

To make a histogram and find summary statistics:

- ▶ Choose **Distribution** from the **Analyze** menu.
- ▶ In the **Distribution** dialog, drag the name of the variable that you wish to analyze into the empty window beside the label "**Y, Columns.**"

- ▶ Click **OK**. JMP computes standard summary statistics along with displays of the variables.

**MINITAB**

To make a histogram:

- ▶ Choose **Histogram** from the **Graph** menu.
- ▶ Select “Simple” for the type of graph and click **OK**.
- ▶ Enter the name of the quantitative variable you wish to display in the box labeled “Graph variables.” Click **OK**.

To calculate summary statistics:

- ▶ Choose **Basic statistics** from the **Stat** menu. From the **Basic Statistics** submenu, choose **Display Descriptive Statistics**.
- ▶ Assign variables from the variable list box to the Variables box. MINITAB makes a Descriptive Statistics table.

**SPSS**

To make a histogram in SPSS open the Chart Builder from the Graphs menu.

- ▶ Click the **Gallery** tab.
- ▶ Choose **Histogram** from the list of chart types.
- ▶ Drag the histogram onto the canvas.
- ▶ Drag a scale variable to the y-axis drop zone.
- ▶ Click **OK**.

To calculate summary statistics:

- ▶ Choose **Explore** from the **Descriptive Statistics** submenu of the **Analyze** menu. In the Explore dialog, assign one or more variables from the source list to the Dependent List and click the **OK** button.

**TI-NSPIRE**

To plot a histogram using a named list, press  $\blacktriangle$  several times so that the entire list is highlighted. Press  $\text{MENU}$ ,  $\text{3}$  for Data, and  $\text{4}$  for Quick Graph. Then press  $\text{MENU}$ ,  $\text{1}$  for Plot Type, and  $\text{3}$  for Histogram.

To create the plot on a full page, press  $\text{GRAPH}$ , and then  $\text{5}$  for Data & Statistics. Move the cursor to “Click to add variable,” and then press  $\text{PAGE DOWN}$  and select the list name. Then press  $\text{MENU}$ ,  $\text{1}$  for Plot Type, and  $\text{3}$  for Histogram.

**TI-89**

To make a histogram:

- ▶ Select  $\text{F2}$  (**Plots**), then 1: **Plot Setup**. Select a plot and press  $\text{F1}$  to define it.
- ▶ Select plot type 4: **Histogram**. Use VAR-LINK to select the data list.
- ▶ Enter a number for the histogram bucket (bar) width.
- ▶ Press  $\text{ENTER}$  to complete the plot definition. Press  $\text{F5}$  to display the histogram.
- ▶ Press  $\blacktriangle\text{F2}$  to adjust the window appropriately, then press  $\blacktriangle\text{F3}$  (**Graph**).

To calculate summary statistics:

- ▶ To compute summary statistics, press  $\text{F4}$  (**Calc**). Input the name of the list using VAR-LINK. Press  $\text{ENTER}$ .
- ▶ Use the down arrow to scroll through the output.
- ▶ To create a boxplot, press  $\text{F2}$  (**Plots**) then  $\text{ENTER}$ . Select a plot to define and press  $\text{F1}$ . Select either 3: **Box Plot** or 4: **Mod Box**

**Plot** (to identify outliers). Select the mark type of your choice (for outliers). Press  $\text{ENTER}$  to finish.

- ▶ Press  $\text{F5}$  to display the graph.

**COMMENTS**

If the data are stored as a frequency table (say, with data values in list1 and frequencies in list2), change Use Freq and Categories to YES and use VAR-LINK to select list2 as the frequency variable on the plot definition screen.

If the data are stored as a frequency table (say, with data values in list1 and frequencies in list2), use VAR-LINK to select list2 as the frequency variable in 1-Var Stats.

For the plot, change Use Freq and Categories to YES and use VAR-LINK to select list2 as the frequency variable on the plot definition screen.

**Chapter 5. Understanding and Comparing Distributions**

There are two ways to organize data when we want to compare groups. Each group can be in its own variable (or list, on a calculator). In this form, the experiment comparing coffee cups would have four lists, one for each type of cup:

CUPPS	SIGG	Nissan	Starbucks
6	2	12	13
6	1.5	16	7
6	2	9	7
18.5	3	23	17.5
10	0	11	10
17.5	7	20.5	15.5
11	0.5	12.5	6
6.5	6	24.5	6

But there's another way to think about and organize the data. What is the variable of interest (the *What*) in this experiment? It's the number of degrees lost by the water in each cup. And the *Who* is each time she tested a cup. We could gather all the temperature values into one variable and put the names of the cups in a second variable listing the individual results, one on each row. Now the *Who* is clearer—it's an experimental run, one row of the table. Most statistics packages prefer data on groups organized in this way.

That's actually the way we've thought about the wind speed data in this chapter, treating wind speeds as one variable and the groups (whether seasons, months, or days) as a second variable.

Container	Temperature Difference	Container	Temperature Difference
CUPPS	6	SIGG	12
CUPPS	6	SIGG	16
CUPPS	6	SIGG	9
.	.	.	.
.	.	.	.
Nissan	2	Starbucks	13
Nissan	1.5	Starbucks	7
Nissan	2	Starbucks	7
.	.	.	.
.	.	.	.
.	.	.	.

## DATA DESK

If the data are in separate variables, select the variables and choose **Boxplot side by side** from the **Plot** menu. The boxes will appear in the order in which the variables were selected. If the data are a single quantitative variable and a second variable holding group names, select the quantitative variable as

Y and the group variable as X. Then choose **Boxplot y by x** from the **Plot** menu. The boxes will appear in alphabetical order by group name. Data Desk offers options for assessing whether any pair of medians differ.

## EXCEL

Excel cannot make boxplots.

## COMMENT

The DDXL add-on provided on the DVD adds the ability to make boxplots to Excel.

## JMP

Choose **Fit y by x**. Assign a continuous response variable to **Y, Response** and a nominal group variable holding the group names to **X, Factor**, and click **OK**. JMP will offer (among other

things) dotplots of the data. Click the red triangle and, under **Display Options**, select **Boxplots**. *Note:* If the variables are of the wrong type, the display options might not offer boxplots.

## MINITAB

Choose **Boxplot...** from the **Graph** menu. If your data are in the form of one quantitative variable and one group variable, choose

**One Y and with Groups**. If your data are in separate columns of the worksheet, choose **Multiple Y's**.

## SPSS

To make a boxplot in SPSS, open the **Chart Builder** from the **Graphs** menu.

Click the **Gallery** tab.

Choose **Boxplot** from the list of chart types.

Drag a single or 2-D (side-by-side) boxplot onto the canvas.

Drag a scale variable to the y-axis drop zone.

To make side-by-side boxplots, drag a categorical variable to the x-axis drop zone.

Click **OK**.

## TI-NSPIRE

To compute summary statistics using a named list, press  $\text{2ND}$ ,  $\text{1}$  for Calculator,  $\text{MENU}$ ,  $\text{6}$  for Statistics,  $\text{1}$  for Stat Calculations, and  $\text{1}$  for One-Variable Statistics. Complete the dialog boxes.

To create a box plot using a named list, press  $\blacktriangle$  several times so that the entire list is highlighted. Press  $\text{MENU}$ ,  $\text{3}$  for Data, and

$\text{4}$  for Quick Graph. Then press  $\text{MENU}$ ,  $\text{1}$  for Plot Type, and  $\text{2}$  for Box Plot.

To create the plot on a full page, press  $\text{2ND}$ , and then press  $\text{5}$  for Data & Statistics. Move the cursor to "Click to add variable," and then press  $\text{F6}$  and select the list name. Then press  $\text{MENU}$ ,  $\text{1}$  for Plot Type, and  $\text{2}$  for Box Plot.

**TI-89**

For the plot, change Use Freq and Categories to YES and use VAR-LINK to select list2 as the frequency variable on the plot definition screen.

To create a boxplot, press **F2** (**Plots**), then **ENTER**. Select a plot to define and press **F1**. Select either 3: **Box Plot** or 4: **Mod Box**

**Plot** (to identify outliers). Select the mark type of your choice (for outliers). Press **ENTER** to finish.

Press **F5** to display the graph.

## Chapter 6. The Standard Deviation as a Ruler and the Normal Model

**DATA DESK**

To make a “Normal Probability Plot” in Data Desk,

- ▶ Select the Variable.
- ▶ Choose **Normal Prob Plot** from the **Plot** menu.

**COMMENTS**

Data Desk places the ordered data values on the vertical axis and the Normal scores on the horizontal axis.

**EXCEL**

Excel offers a “Normal probability plot” as part of the Regression command in the Data Analysis extension, but (as of this writing)

it is not a correct Normal probability plot and should not be used.

**JMP**

To make a “Normal Quantile Plot” in JMP,

- ▶ Make a histogram using **Distributions** from the **Analyze** menu.
- ▶ Click on the drop-down menu next to the variable name.
- ▶ Choose **Normal Quantile Plot** from the drop-down menu.
- ▶ JMP opens the plot next to the histogram.

**COMMENTS**

JMP places the ordered data on the vertical axis and the Normal scores on the horizontal axis. The vertical axis aligns with the histogram’s axis, a useful feature.

**MINITAB**

To make a “Normal Probability Plot” in MINITAB,

- ▶ Choose **Probability Plot** from the **Graph** menu.
- ▶ Select “Single” for the type of plot. Click **OK**.
- ▶ Enter the name of the variable in the “Graph variables” box. Click **OK**.

**COMMENTS**

MINITAB places the ordered data on the horizontal axis and the Normal scores on the vertical axis.

**SPSS**

To make a Normal “P-P plot” in SPSS,

- ▶ Choose **P-P** from the **Graphs** menu.
- ▶ Select the variable to be displayed in the source list.
- ▶ Click the arrow button to move the variable into the target list.
- ▶ Click the **OK** button.

**COMMENTS**

SPSS places the ordered data on the horizontal axis and the Normal scores on the vertical axis. You may safely ignore the options in the P-P dialog.

**TI-NSPIRE**

To create a normal probability plot using a named list, press **▲** several times so that the entire list is highlighted. Press **menu**, **3** for Data, and **4** for Quick Graph. Then press **menu**, **1** for Plot Type, and **4** for Normal Probability Plot.

To create the plot on a full page, press **2nd**, and then **5** for Data & Statistics. Move the cursor to “Click to add variable,” and then press **2nd** and select the list name. Then press **menu**, **1** for Plot Type, and **4** for Normal Probability Plot.

To compute the area under a normal curve, press **menu**, **1** for Calculator, **menu**, **5** for Probability, **5** for Distributions, and **2** for Normal Cdf. Complete the dialog box.

To compute the value for a given percentile, press **menu**, **1** for Calculator, **menu**, **5** for Probability, **5** for Distributions, **3** for Inverse Normal. Complete the dialog box.

## TI-89

- ▶ To create a “Normal Prob Plot”, press **F2** and select choice 2: **Norm Prob Plot**. Select a plot number and use VAR-LINK to enter the data list. Select X or Y for the data axis. Press **ENTER** to calculate the z-scores.
- ▶ Press **F2** and select choice 1: **Plot Setup**. Turn off any undesired plots (either **F3** (Clear) or **F4** (✓)). Press **F5** to display the plot.
- ▶ To find what percent of a Normal model lies between two z-scores, press **F5** (**Distr**). Then select 4: **Normal Cdf**. Enter the lower and upper z-scores, specify mean 0 and standard deviation 1, and press **ENTER**.
- ▶ To find the z-score for a given percentile, press **F5** (**Distr**). Then arrow down to 2: **Inverse** press the right arrow to see the sub

menu and select 1: **Inverse Normal**. Enter the area to the left of the desired point, mean 0 and standard deviation 1, and press **ENTER**.

## COMMENTS

Normal models strictly go to infinity on either end, which is 1EE99 on the calculator. In practice, any “large” number will work. For example, the percentage of the Normal model over two standard deviations above the mean can use Lower Value 2 and Upper Value 99. To find area more than 2 standard deviations below the mean, use Lower Value  $-99$ , and Upper value  $-2$ .

## Chapter 7. Scatterplots, Association, and Correlation

## DATA DESK

To make a scatterplot of two variables, select one variable as Y and the other as X and choose **Scatterplot** from the **Plot** menu. Then find the correlation by choosing **Correlation** from the scatterplot’s HyperView menu.

Alternatively, select the two variables and choose **Pearson Product-Moment** from the **Correlations** submenu of the **Calc** menu.

## COMMENTS

We prefer that you look at the scatterplot first and then find the correlation. But if you’ve found the correlation first, click on the correlation value to drop down a menu that offers to make the scatterplot.

## EXCEL

To make a Scatterplot with the Excel Chart Wizard:

- ▶ Click on the **Chart Wizard** Button in the menu bar. Excel opens the Chart Wizard’s Chart Type Dialog window.
- ▶ Make sure the **Standard Types** tab is selected, and select **XY (Scatter)** from the choices offered.
- ▶ Specify the **scatterplot without lines** from the choices offered in the Chart subtype selections. The **Next** button takes you to the Chart Source Data dialog.
- ▶ If it is not already frontmost, click on the **Data Range** tab, and enter the data range in the space provided.
- ▶ By convention, we always represent variables in columns. The Chart Wizard refers to variables as Series. Be sure the **Column** option is selected.
- ▶ Excel places the leftmost column of those you select on the x-axis of the scatterplot. If the column you wish to see on the x-axis is not the leftmost column in your spreadsheet, click on the **Series** tab and edit the specification of the individual axis series.
- ▶ Click the **Next** button. The Chart Options dialog appears.
- ▶ Select the **Titles** tab. Here you specify the title of the chart and names of the variables displayed on each axis.
- ▶ Type the chart title in the **Chart title:** edit box.
- ▶ Type the x-axis variable name in the **Value (X) Axis:** edit box. Note that you must name the columns correctly here. Naming another variable will not alter the plot, only mislabel it.

- ▶ Type the y-axis variable name in the **Value (Y) Axis:** edit box.
- ▶ Click the **Next** button to open the chart location dialog.
- ▶ Select the **As new sheet:** option button.
- ▶ Click the **Finish** button.

Often, the resulting scatterplot will not be useful. By default, Excel includes the origin in the plot even when the data are far from zero. You can adjust the axis scales.

To change the scale of a plot axis in Excel:

- ▶ Double-click on the axis. The **Format Axis Dialog** appears.
- ▶ If the **scale tab** is not the frontmost, select it.
- ▶ Enter new minimum or new maximum values in the spaces provided. You can drag the dialog box over the scatterplot as a straightedge to help you read the maximum and minimum values on the axes.
- ▶ Click the **OK** button to view the rescaled scatterplot.
- ▶ Follow the same steps for the x-axis scale.

Compute a correlation in Excel with the **CORREL** function from the drop-down menu of functions. If **CORREL** is not on the menu, choose **More Functions** and find it among the statistical functions in the browser.

In the dialog that pops up, enter the range of cells holding one of the variables in the space provided.

Enter the range of cells for the other variable in the space provided.



## EXCEL 2007

To make a scatterplot in Excel 2007:

- ▶ Select the columns of data to use in the scatterplot. You can select more than one column by holding down the control key while clicking.
- ▶ In the Insert tab, click on the **Scatter** button and select the **Scatter with only Markers** chart from the menu.

Unfortunately, the plot this creates is often statistically useless.

To make the plot useful, we need to change the display:

- ▶ With the chart selected click on the **Gridlines** button in the Layout tab to cause the Chart Tools tab to appear.
- ▶ Within Primary Horizontal Gridlines, select **None**. This will remove the gridlines from the scatterplot.
- ▶ To change the axis scaling, click on the numbers of each axis of the chart, and click on the **Format Selection** button in the Layout tab.
- ▶ Select the **Fixed** option instead of the Auto option, and type a value more suited for the scatterplot. You can use the popup dialog window as a straightedge to approximate the appropriate values.

Excel 2007 automatically places the leftmost of the two columns you select on the x-axis, and the rightmost one on the y-axis. If that's not what you'd prefer for your plot, you'll want to switch them.

To switch the X and Y-variables:

- ▶ Click the chart to access the **Chart Tools** tabs.
- ▶ Click on the **Select Data** button in the Design tab.
- ▶ In the popup window's Legend Entries box, click on **Edit**.
- ▶ Highlight and delete everything in the Series X Values line, and select new data from the spreadsheet. (Note that selecting the column would inadvertently select the title of the column, which would not work well here.)
- ▶ Do the same with the Series Y Values line.
- ▶ Press **OK**, then press **OK** again.

## JMP

To make a scatterplot and compute correlation, choose **Fit Y by X** from the **Analyze** menu.

In the Fit Y by X dialog, drag the Y variable into the "**Y, Response**" box, and drag the X variable into the "**X, Factor**" box. Click the **OK** button.

Once JMP has made the scatterplot, click on the red triangle next to the plot title to reveal a menu of options. Select **Density Ellipse** and select .95. JMP draws an ellipse around the data and reveals the **Correlation** tab. Click the blue triangle next to Correlation to reveal a table containing the correlation coefficient.

## MINITAB

To make a scatterplot, choose **Scatterplot** from the **Graph** menu. Choose "Simple" for the type of graph. Click **OK**. Enter variable names for the Y-variable and X-variable into the table. Click **OK**.

To compute a correlation coefficient, choose **Basic Statistics** from the **Stat** menu. From the Basic Statistics submenu, choose **Correlation**. Specify the names of at least two quantitative variables in the "Variables" box. Click **OK** to compute the correlation table.

## SPSS

To make a scatterplot in SPSS, open the Chart Builder from the Graphs menu. Then:

- ▶ Click the Gallery tab.
- ▶ Choose Scatterplot from the list of chart types.
- ▶ Drag the scatterplot onto the canvas.
- ▶ Drag a scale variable you want as the response variable to the y-axis drop zone.
- ▶ Drag a scale variable you want as the factor or predictor to the x-axis drop zone.
- ▶ Click OK.

To compute a correlation coefficient, choose **Correlate** from the **Analyze** menu. From the Correlate submenu, choose **Bivariate**. In the Bivariate Correlations dialog, use the arrow button to move variables between the source and target lists. Make sure the **Pearson** option is selected in the Correlation Coefficients field.

## TI-NSPIRE

To create a scatterplot using named lists, press  $\blacktriangle$  several times so that the first list is highlighted. Then press  $\leftarrow$  so that the second list is highlighted. Press  $\left[\text{menu}\right]$ ,  $\left[3\right]$  for Data, and  $\left[4\right]$  for Quick Graph.

To create the plot on a full page, press  $\left[\text{fn}\right]$ , then  $\left[5\right]$  for Data & Statistics. Move the cursor to "Click to add variable," and

then press  $\left[\text{2nd}\right]$  and select the list name. Repeat for the other axis.

To find the correlation, press  $\left[\text{fn}\right]$ ,  $\left[1\right]$  for Calculator,  $\left[\text{menu}\right]$ ,  $\left[6\right]$  for Statistics,  $\left[1\right]$  for Stat Calculations, and  $\left[4\right]$  for Linear Regression. Complete the dialog boxes.

## TI-89

To create a scatterplot, press **F2** (**Plots**). Select choice 1: **Plot Setup**. Select a plot to define and press **F1**. Select **Plot Type 1: Scatter**. Select a mark type. Specify the lists where the data are stored as Xlist and Ylist, using VAR-LINK. Press **ENTER** to finish. Press **F5** to display the plot.

To find the correlation, press **F4** (**CALC**), then arrow to **3: Regressions**, press the right arrow, and select **1: LinReg(a+bx)**.

Then specify the lists where the data are stored. You can also select a y-function to store the equation of the line.

## COMMENTS

Notice that if you **TRACE** (press **F3**) the scatterplot, the calculator will tell you the x- and y-value at each point.

## Chapter 8. Linear Regression

## DATA DESK

Select the y-variable and the x-variable. In the **Plot** menu choose **Scatterplot**. from the scatterplot HyperView menu, choose **Add Regression Line** to display the line. from the HyperView menu, choose **Regression** to compute the regression.

## COMMENTS

Alternatively, find the regression first with the **Regression** command in the **Calc** menu. Click on the x-variable's name to open a menu that offers the scatterplot.

## EXCEL

Make a scatterplot of the data. With the scatterplot front-most, select **Add Trendline...** from the **Chart** menu. Click the **Options** tab and select **Display Equation on Chart**. Click **OK**.

## COMMENTS

The computer section for Chapter 7 shows how to make a scatterplot. We don't repeat those steps here.

## EXCEL 2007

- ▶ Click on a blank cell in the spreadsheet.
- ▶ Go to the **Formulas** tab in the Ribbon and click **More Functions** → **Statistical**.
- ▶ Choose the **CORREL** function from the drop-down menu of functions.
- ▶ In the dialog that pops up, enter the range of one of the variables in the space provided.
- ▶ Enter the range of the other variable in the space provided.
- ▶ Click **OK**.

## COMMENTS

The correlation is computed in the selected cell. Correlations computed this way will update if any of the data values are changed. Before you interpret a correlation coefficient, always make a scatterplot to check for nonlinearity and outliers. If the variables are not linearly related, the correlation coefficient cannot be interpreted.

## JMP

Choose **Fit Y by X** from the **Analyze** menu. Specify the y-variable in the Select Columns box and click the **"Y, Response"** button. Specify the x-variable and click the **"X, Factor"** button. Click **OK** to make a scatterplot. In the

scatterplot window, click on the red triangle beside the heading labeled "Bivariate Fit . . ." and choose **"Fit Line."** JMP draws the least squares regression line on the scatterplot and displays the results of the regression in tables below the plot.

## MINITAB

Choose **Regression** from the **Stat** menu. From the Regression submenu, choose **Fitted Line Plot**. In the Fitted Line Plot dialog, click in the **Response Y** box, and assign the y-variable from the

Variable list. Click in the **Predictor X** box, and assign the x-variable from the Variable list. Make sure that the Type of Regression Model is set to Linear. Click the **OK** button.

## SPSS

Choose **Interactive** from the **Graphs** menu. From the interactive Graphs submenu, choose **Scatterplot**. In the Create Scatterplot dialog, drag the y-variable into the **y-axis target**, and the

x-variable into the **x-axis target**. Click on the **Fit** tab. Choose **Regression** from the **Method** popup menu. Click the **OK** button.

**TI-NSPIRE**

To plot and find the equation of the regression line, first create a scatterplot. Using named lists, press  $\blacktriangle$  several times so that the first list is highlighted. Then press  $\langle \text{2} \rangle$  so that the second list is highlighted. Press  $\langle \text{menu} \rangle$ ,  $\langle \text{3} \rangle$  for Data, and  $\langle \text{4} \rangle$  for Quick Graph. Then press  $\langle \text{menu} \rangle$ ,  $\langle \text{3} \rangle$  for Actions,  $\langle \text{5} \rangle$  for Regression, and  $\langle \text{2} \rangle$  for Show Linear.

To find the equation of the regression line on a full page, press  $\langle \text{2nd} \rangle$ ,  $\langle \text{1} \rangle$  for Calculator,  $\langle \text{menu} \rangle$ ,  $\langle \text{6} \rangle$  for Statistics,  $\langle \text{1} \rangle$  for Stat Calcu-

lations, and  $\langle \text{4} \rangle$  for Linear Regression. Complete the dialog boxes.

To see the plot on a full page, press  $\langle \text{2nd} \rangle$ , and then  $\langle \text{5} \rangle$  for Data & Statistics. Move the cursor to “Click to add variable,” and then press  $\langle \text{2nd} \rangle$  and select the list name. Repeat for the other axis. Then press  $\langle \text{menu} \rangle$ ,  $\langle \text{3} \rangle$  for Actions,  $\langle \text{5} \rangle$  for Regression, and  $\langle \text{2} \rangle$  for Show Linear.

**TI-89**

To find the equation of the regression line (and add the line to a scatterplot), choose **LinReg (a+bx)** from the **Calc Regressions** menu and tell it the list names and a function to store the equation. To make a residuals plot, define a **PLOT** as a scatterplot. Specify your explanatory datalist as Xlist. For Ylist, find the list name **resid** from VAR-LINK by arrowing to the **STATVARS** portion. then press  $\langle \text{2} \rangle$  (**r**) and locate the list. press  $\langle \text{ENTER} \rangle$  to finish the plot definition and  $\langle \text{F5} \rangle$  to display the plot.

**COMMENTS**

Each time you execute a **LinReg** command, the calculator automatically computes the residuals and stores them in a data list named RESID. If you don't want to see this (or any other calculator-generated list) anymore, press  $\langle \text{F1} \rangle$  (Tools) and select choice 3: Setup Editor. Leaving the box for lists to display blank will reset the calculator to show only lists 1 through 6.

**Chapter 9. Regression Wisdom****DATA DESK**

Click on the **HyperView** menu on the **Regression** output table. A menu drops down to offer scatterplots of residuals against predicted values, Normal probability plots of residuals, or just the ability to save the residuals and predicted values. Click on the name of a predictor in the regression table to be offered a scatterplot of the residuals against that predictor.

**COMMENTS**

If you change any of the variables in the regression analysis, Data Desk will offer to update the plots of residuals.

**EXCEL**

The Data Analysis add-in for Excel includes a Regression command. The dialog box it shows offers to make plots of residuals.

**COMMENTS**

Do not use the Normal probability plot offered in the regression dialog. It is not what it claims to be and is wrong.

**JMP**

From the **Analyze** menu, choose **Fit Y by X**. Select **Fit Line**. Under Linear Fit, Select **Plot Residuals**. You can also choose

to **Save Residuals**. Subsequently, from the **Distribution** menu, choose **Normal quantile plot** or **histogram** for the residuals.

**MINITAB**

From the **Stat** menu, choose **Regression**. From the **Regression** submenu, select **Regression** again. In the Regression dialog, enter the response variable name in the “Response” box and the predictor variable name in the “Predictor” box. To specify saved results, in the Regression dialog, click **Storage**. Check “Residu-

als” and “Fits.” Click **OK**. To specify displays, in the Regression dialog, click **Graphs**. Under “Residual Plots,” select “Individual plots” and check “Residuals versus fits.” Click **OK**. Now back in the Regression dialog, click **OK**. Minitab computes the regression and the requested saved values and graphs.

**SPSS**

From the **Analyze** menu, choose **Regression**. From the Regression submenu, choose **Linear**. After assigning variables to their roles in the regression, click the “**Plots...**” button. In the Plots dialog, you can specify a Normal probability plot of residuals and scatterplots of various versions of standardized residuals and predicted values.

**COMMENTS**

A plot of **\*ZRESID** against **\*PRED** will look most like the residual plots we've discussed. SPSS standardizes the residuals by dividing by their standard deviation. (There's no need to subtract their mean; it must be zero.) The standardization doesn't affect the scatterplot.

**TI-NSPIRE**

To create a residual plot, press  $\langle \text{2nd} \rangle$ , then  $\langle \text{5} \rangle$  for Data & Statistics. Move the cursor to “Click to add variable,” and then press  $\langle \text{2nd} \rangle$

and select the list name. For the other axis, select the variable name **stat.resid**.

### TI-89

To make a residuals plot, define a Plot as a scatterplot. Specify your explanatory datalist as **Xlist**. For **Ylist**, find the list name resid from **VAR-LINK** by arrowing to the **STATVARS** portion. Then press **[2]** (**r**) and locate the list. Press **[ENTER]** to finish the plot definition and **[F5]** to display the plot.

### COMMENTS

Each time you execute a **LinReg** command, the calculator automatically computes the residuals and stores them in a data list named **RESID**. If you don't want to see this (or any other calculator-generated list) anymore, press **[F1]** (Tools) and select choice 3: Setup Editor. Leaving the box for lists to display blank will reset the calculator to show only lists 1 through 6.

## Chapter 10. Re-expressing Data: Get It Straight!

### DATA DESK

To re-express a variable in Data Desk, select the variable and Choose the function to re-express it from the **Manip > Transform** menu. Square root, log, reciprocal, and reciprocal root are immediately available. For others, make a derived variable and type the function. Data Desk makes a new derived variable that holds the re-expressed values. Any value changed in the original variable will immediately be re-expressed in the derived variable.

### COMMENTS

Or choose **Manip > Transform > Dynamic > Box-Cox** to generate a continuously changeable variable and a slider that specifies the power. Set plots to **Automatic Update** in their HyperView menus and watch them change dynamically as you drag the slider.

### EXCEL

To re-express a variable in Excel, use Excel's built-in functions as you would for any calculation. Changing a value in the original column will change the re-expressed value.

### JMP

To re-express a variable in JMP, double-click to the right of the last column of data to create a new column. Name the new column and select it. Choose **Formula** from the **Cols** menu. In the Formula dialog, choose the transformation and variable that you wish to assign to the new column. Click the **OK** button. JMP places the re-expressed data in the new column.

### COMMENTS

The log and square root re-expressions are found in the **Transcendental** menu of functions in the formula dialog.

### MINITAB

To re-express a variable in MINITAB, choose **Calculator** from the **Calc** menu. In the Calculator dialog, specify a name for the new re-expressed variable. Use the **Functions List**, the calculator

buttons, and the **Variables list** box to build the expression. Click **OK**.

### SPSS

To re-express a variable in SPSS, Choose **Compute** from the **Transform** menu. Enter a name in the Target Variable field. Use the calculator and Function List to build the expression. Move a

variable to be re-expressed from the source list to the Numeric Expression field. Click the **OK** button.

### TI-NSPIRE

To re-express data, create a new list and enter the formula in the cell in the second row. For example, if one column has a list

named *time*, another list can be created using the formula  $\log(\text{time})$ .

### TI-89

To re-express data stored in a list, perform the re-expression on the whole list and store it in another list. For example, to use the common (base 10) logarithms of the data in list1, on the home screen, enter the command **log(list1)** **[STO]** **list2**.

### COMMENTS

- ▶ To find the log command, press **[CATALOG]** then **[4]** (L) arrow to log, and press **[ENTER]**.
- ▶ Natural logs are **LN** (press **[2nd][X]**).
- ▶ For square roots, press **[2nd][√]**.

## Chapter 11. Understanding Randomness

<b>DATA DESK</b>	<p>Generate random numbers in Data Desk with the <b>Generate Random Numbers . . .</b> command in the <b>Manip</b> menu. A dialog guides you in specifying the number of variables to fill, the number of cases, and details about the values. For most simulations, generate random uniform values.</p>	<p><b>COMMENTS</b></p> <p><b>Bernoulli Trials</b> generate random values that are 0 or 1, with a specified chance of a 1.</p> <p><b>Binomial Experiments</b> automatically generate a specified number of Bernoulli trials and count the number of 1's.</p>
<b>EXCEL</b>	<p>The <b>RAND</b> function generates a random value between 0 and 1. You can multiply to scale it up to any range you like and use the <b>INT</b> function to turn the result into an integer.</p>	<p><b>COMMENTS</b></p> <p>Published tests of Excel's random-number generation have declared it to be inadequate. However, for simple simulations, it should be OK. Don't trust it for important large simulations.</p>
<b>JMP</b>	<p>In a new column, in the <b>Cols</b> menu choose <b>Column Info...</b> In the dialog, click the <b>New Property</b> button, and choose <b>Formula</b> from the drop-down menu.</p>	<p>Click the <b>Edit Formula</b> button, and in the <b>Functions(grouped)</b> window click on <b>Random. Random Integer (10)</b>, for example, will generate a random integer between 1 and 10.</p>
<b>MINITAB</b>	<p>In the <b>Calc</b> menu, choose <b>Random Data . . .</b> In the Random Data submenu, choose <b>Uniform . . .</b></p>	<p>A dialog guides you in specifying details of range and number of columns to generate.</p>
<b>SPSS</b>	<p>The <b>RV.UNIFORM(min, max)</b> function returns a random value that is equally likely between the min and max limits.</p>	
<b>TI-NSPIRE</b>	<p>To generate random integers, press <math>\text{ⓐ}</math>, <math>\text{ⓑ}</math> for Calculator, <math>\text{ⓓ}</math> for Probability, <math>\text{Ⓔ}</math> for Random, and <math>\text{Ⓒ}</math> for Integer. Then type the range for the random integers, such as <code>randInt(1,6)</code>.</p>	<p>To create a list of random integers, type the length of the list as the third value, such as <code>randInt(1,6,10)</code>.</p>
<b>TI-89</b>	<p>To generate random numbers, move the cursor to highlight the name of a blank list. Use <b>5:RandInt</b> from the <math>\text{ⓕ}</math> (Calc) Probability menu. This command will produce any number of random integers in the specified range.</p>	<p><b>COMMENTS</b></p> <p>Some examples:</p> <p><b>RandInt(0,10)</b> randomly chooses a 0 or a 1. This is an effective simulation of 10 coin tosses.</p> <p><b>RandInt(1,6,2)</b> randomly returns two integers between 1 and 6. This is a good way to simulate rolling two dice.</p> <p><b>RandInt(0,56,3)</b> produces three random integers between 0 and 56, a nice way to simulate the chapter's dorm room lottery.</p>

## Chapter 16. Random Variables

<b>TI-NSPIRE</b>	<p>To compute the mean and standard deviation for a discrete random variable, enter the values in one named list and the probabilities in another. Then press <math>\text{ⓐ}</math>, <math>\text{ⓑ}</math> for Calculator, <math>\text{ⓓ}</math> for Statistics, <math>\text{ⓑ}</math> for Stat Calculations, and <math>\text{ⓑ}</math> for One-Variable Statistics. Enter 2 for the prompt for the number of lists, <math>\text{ⓐ}</math> to OK, <math>\text{ⓐ}</math>, and complete the dialog box.</p>	
<b>TI-89</b>	<p>To calculate the mean and standard deviation of a discrete random variable, enter the probability model in two lists:</p> <ul style="list-style-type: none"> <li>▶ In one list (say, list1) enter the <math>x</math>-values of the variable.</li> <li>▶ In a second list (say, list2) enter the associated probabilities <math>P(X = x)</math>.</li> <li>▶ From the <b>STAT CALC</b> (<math>\text{ⓕ}</math>) menu select <b>1-VarStats</b>. Use <b>VAR-LINK</b> to enter the list name list1 in the List box and list2 in the Freq box.</li> </ul>	<p><b>COMMENTS</b></p> <p>You can enter the probabilities as fractions; the calculator will change them to decimals for you.</p> <p>Notice that the calculator knows enough to compute only the standard deviation <math>\sigma</math>, but mistakenly uses <math>\bar{x}</math> when it should say <math>\mu</math>. Make sure you don't make that mistake!</p>

## Chapter 17. Probability Models

The only important differences among these functions are in what they are named and the order of their arguments. In these functions, pdf stands for “probability density function”—what we’ve been calling a probability model. The letters cdf stand for “cumulative distribution function,” the technical term when we want to accumulate probabilities over a range of values. These technical terms show up in many of the function names. The term “cumulative” in a function name says that it corresponds to a cdf.

Generically, the four functions are as follows:

Geometric pdf ( <i>prob</i> , <i>x</i> )	Finds the individual geometric probability of getting the first success on trial <i>x</i> when the probability of success is <i>prob</i> .	For example, the probability of finding the first Tiger Woods picture in the fifth cereal box is Geometric pdf(0.2, 5)
Geometric cdf ( <i>prob</i> , <i>x</i> )	Finds the cumulative probability of getting the first success on or before trial <i>x</i> , when the probability of success is <i>prob</i> .	For example, the total probability of finding Tiger’s picture in one of the first 4 boxes is Geometric cdf(0.2, 4)
Binomial pdf ( <i>n</i> , <i>prob</i> , <i>x</i> )	Finds the probability of getting <i>x</i> successes in <i>n</i> trials when the probability of success is <i>prob</i> .	For example, Binomial pdf(5, 0.2, 2) is the probability of finding Tiger’s picture exactly twice among 5 boxes of cereal.

### DATA DESK

**BinomDistr**(*x*, *n*, *prob*) (pdf)  
**CumBinomDistr**(*x*, *n*, *prob*) (cdf)

### COMMENTS

Data Desk does not compute Geometric probabilities. These functions work in derived variables or in scratchpads.

### EXCEL

**Binomdist**(*x*, *n*, *prob*, *cumulative*)

### COMMENTS

Set *cumulative* = *true* or for cdf, *false* for pdf. Excel’s function fails when *x* or *n* is large. Possibly, it does not use the Normal approximation. Excel does not compute Geometric probabilities.

### JMP

**Binomial Probability** (*prob*, *n*, *x*) (pdf)  
**Binomial Distribution** (*prob*, *n*, *x*) (cdf)

### COMMENTS

JMP does not compute Geometric probabilities.

### MINITAB

Choose **Probability Distributions** from the **Calc** menu. Choose **Binomial** from the Probability Distributions submenu. To calculate the probability of getting *x* successes in *n* trials, choose **Probability**.

To calculate the probability of getting *x* or fewer successes among *n* trials, choose **Cumulative Probability**. For Geometric, choose **Geometric** from the Probability Distribution submenu.

### SPSS

**PDF.GEOM**(*x*, *prob*)  
**CDF.GEOM**(*x*, *prob*)

**PDF.BINOM**(*x*, *n*, *prob*)  
**CDF.BINOM**(*x*, *n*, *prob*)

### TI-NSPIRE

To compute geometric and binomial probabilities, press  $\left(\text{menu}\right)$ ,  $\left\langle 5 \right\rangle$  for Probability, and  $\left\langle 5 \right\rangle$  for Distributions. Select the menu item.

Pdf is for the probability distribution function; Cdf will display cumulative probabilities. Complete the dialog box.

### TI-89

Find the commands under the  $\left[\text{F5}\right]$  (Distributions) menu.

- ▶ F: **Geometric Pdf** will ask for *p* and *x*. It returns the probability of the first success occurring on the *x*th trial.
- ▶ G: **Geometric Cdf** will ask for *p* and the upper and lower values of interest, say *a* and *b*. It returns  $P(a \leq X \leq b)$ , the probability the first success occurs between the *a*th and *b*th trials, inclusive.
- ▶ A: **Binomial Pdf** asks for *n*, *p*, and *x*.
- ▶ B: **Binomial Cdf** asks for *n*, *p*, and the lower and upper values of interest.

### COMMENTS

For Geometric variables, when finding  $P(X \geq a)$  specify an upper value of infinity,  $1\left[\text{EE}\right]99$ , or a very large number. For Binomial variables, when finding  $P(X \geq a)$ , the upper value is *n*.

## Chapter 19. Confidence Intervals for Proportions

<p><b>DATA DESK</b></p> <p>Data Desk does not offer built-in methods for inference with proportions.</p>	<p><b>COMMENTS</b></p> <p>For summarized data, open a Scratchpad to compute the standard deviation and margin of error by typing the calculation. Then use <b>z-interval for individual <math>\mu</math>s</b>.</p>
<p><b>EXCEL</b></p> <p>Inference methods for proportions are not part of the standard Excel tool set.</p>	<p><b>COMMENTS</b></p> <p>For summarized data, type the calculation into any cell and evaluate it.</p>
<p><b>JMP</b></p> <p>For a <b>categorical</b> variable that holds category labels, the <b>Distribution</b> platform includes tests and intervals for proportions. For summarized data, put the category names in one variable and the frequencies in an adjacent variable. Designate the frequency column to have the <b>role of frequency</b>. Then use the <b>Distribution</b> platform.</p>	<p><b>COMMENTS</b></p> <p>JMP uses slightly different methods for proportion inferences than those discussed in this text. Your answers are likely to be slightly different, especially for small samples.</p>
<p><b>MINITAB</b></p> <p>Choose <b>Basic Statistics</b> from the <b>Stat</b> menu.</p> <ul style="list-style-type: none"> <li>▶ Choose <b>1Proportion</b> from the Basic Statistics submenu.</li> <li>▶ If the data are category names in a variable, assign the variable from the variable list box to the <b>Samples in columns</b> box. If you have summarized data, click the <b>Summarized Data</b> button and fill in the number of trials and the number of successes.</li> <li>▶ Click the <b>Options</b> button and specify the remaining details.</li> </ul>	<ul style="list-style-type: none"> <li>▶ If you have a large sample, check <b>Use test and interval based on normal distribution</b>. Click the <b>OK</b> button.</li> </ul> <p><b>COMMENTS</b></p> <p>When working from a variable that names categories, MINITAB treats the last category as the “success” category. You can specify how the categories should be ordered.</p>
<p><b>SPSS</b></p> <p>SPSS does not find confidence intervals for proportions.</p>	
<p><b>TI-NSPIRE</b></p> <p>To compute a confidence interval for a population proportion, press <math>\text{[2nd]}</math>, <math>\text{[1]}</math> for Calculator, <math>\text{[menu]}</math>, <math>\text{[6]}</math> for Statistics, <math>\text{[6]}</math> for Confidence Intervals, and <math>\text{[5]}</math> for 1-Prop z-interval. Complete the</p>	<p>dialog box. Be sure to enter the number of successes, <math>x</math>, as a whole number, and the C level as a decimal, such as .99.</p>
<p><b>TI-89</b></p> <p>To calculate a confidence interval for a population proportion:</p> <ul style="list-style-type: none"> <li>▶ Go to the <b>Ints</b> menu (<math>\text{[2nd]}</math><math>\text{[F2]}</math>) and select <b>5:1-PropZInt</b>.</li> <li>▶ Enter the number of successes observed and the sample size.</li> <li>▶ Specify a confidence level.</li> <li>▶ Calculate the interval.</li> </ul>	<p><b>COMMENTS</b></p> <p><i>Beware:</i> When you enter the value of <math>x</math>, you need the count, not the percentage. The count must be a whole number. If the number of successes are given as a percentage, you must first multiply <math>np</math> and round the result.</p>

## Chapter 20. Testing Hypotheses About Proportions

<p><b>DATA DESK</b></p> <p>Data Desk does not offer built-in methods for inference with proportions. The <b>Replicate Y by X</b> command in the <b>Manip</b> menu will “reconstruct” summarized count data so that you can display it.</p>	<p><b>COMMENTS</b></p> <p>For summarized data, open a Scratchpad to compute the standard deviation and margin of error by typing the calculation. Then perform the test with the <b>z-test for individual <math>\mu</math>s</b> found in the Test command.</p>
<p><b>EXCEL</b></p> <p>Inference methods for proportions are not part of the standard Excel tool set.</p>	<p><b>COMMENTS</b></p> <p>For summarized data, type the calculation into any cell and evaluate it.</p>

**JMP**

For a **categorical** variable that holds category labels, the **Distribution** platform includes tests and intervals of proportions. For summarized data, put the category names in one variable and the frequencies in an adjacent variable. Designate the frequency column to have the **role of frequency**. Then use the **Distribution** platform.

**COMMENTS**

JMP uses slightly different methods for proportion inferences than those discussed in this text. Your answers are likely to be slightly different.

**MINITAB**

Choose **Basic Statistics** from the **Stat** menu.

- ▶ Choose **1Proportion** from the Basic Statistics submenu.
- ▶ If the data are category names in a variable, assign the variable from the variable list box to the **Samples in columns** box.
- ▶ If you have summarized data, click the **Summarized Data** button and fill in the number of trials and the number of successes.
- ▶ Click the **Options** button and specify the remaining details.

- ▶ If you have a large sample, check **Use test and interval based on Normal distribution**.
- ▶ Click the **OK** button.

**COMMENTS**

When working from a variable that names categories, MINITAB treats the last category as the “success” category. You can specify how the categories should be ordered.

**SPSS**

SPSS does not find hypothesis tests for proportions.

**TI-NSPIRE**

To compute a hypothesis test for a population proportion, press  $\left(\frac{\square}{\square}\right)$ ,  $\langle 1 \rangle$  for Calculator,  $\left(\frac{\square}{\square}\right)$ ,  $\langle 6 \rangle$  for Statistics,  $\langle 7 \rangle$  for Stat Tests,

and  $\langle 5 \rangle$  for 1-Prop z-test. Complete the dialog box. Be sure to enter the number of successes,  $x$ , as a whole number.

**TI-89**

To do the mechanics of a hypothesis test for a proportion,

- ▶ Select **5:1-PropZTest** from the **STAT TESTS**  $\left[\frac{2}{nd}\right][F1]$  menu.
- ▶ Specify the hypothesized proportion.
- ▶ Enter the observed value of  $x$ .
- ▶ Specify the sample size.
- ▶ Indicate what kind of test you want: one-tail lower tail, two-tail, or one-tail upper tail.

- ▶ Specify whether to calculate the result or draw the result (a normal curve with  $p$ -value area shaded.)

**COMMENTS**

*Beware:* When you enter the value of  $x$ , you need the *count*, not the percentage. The count must be a whole number. If the number of successes is given as a percent, you must first multiply  $np$  and round the result to obtain  $x$ .

**Chapter 22. Comparing Two Proportions****DATA DESK**

Data Desk does not offer built-in methods for inference with proportions. Use **Replicate Y by X** to construct data corresponding to given proportions and totals.

**COMMENTS**

For summarized data, open a Scratchpad to compute the standard deviations and margin of error by typing the calculation.

**EXCEL**

Inference methods for proportions are not part of the standard Excel tool set.

**COMMENTS**

For summarized data, type the calculation into any cell and evaluate it.

**JMP**

For a **categorical** variable that holds category labels, the **Distribution** platform includes tests and intervals of proportions. For summarized data, put the category names in one variable and the frequencies in an adjacent variable. Designate the frequency column to have the **role of frequency**. Then use the **Distribution** platform.

**COMMENTS**

JMP uses slightly different methods for proportion inferences than those discussed in this text. Your answers are likely to be slightly different.



**MINITAB**

To find a hypothesis test for a proportion, Choose **Basic Statistics** from the **Stat** menu. Choose **2Proportions . . .** from the Basic Statistics submenu. If the data are organized as category names in one column and case IDs in another, assign the variables from the variable list box to the **Samples in one column** box. If the data are organized as two separate columns of responses, click on **Samples in different columns:** and assign the variables from the variable list box. If you have summarized data, click the **Summarized Data** button and fill in the number of trials and the number of successes for each group.

Click the **Options** button and specify the remaining details. Remember to click the **Use pooled estimate of  $p$  for test** box when testing the null hypothesis of no difference between proportions. Click the **OK** button.

**COMMENTS**

When working from a variable that names categories, MINITAB treats the last category as the “success” category. You can specify how the categories should be ordered.

**SPSS**

SPSS does not find hypothesis tests for proportions.

**TI-NSPIRE**

To compute a confidence interval for the difference between two population proportions, press  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Calculator,  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Statistics,  $\left[\frac{\square}{\square}\right]$  for Confidence Intervals, and  $\left[\frac{\square}{\square}\right]$  for 2-Prop z-interval. Complete the dialog box. Be sure to enter each number of successes as a whole number, and the C level as a decimal, such as .99.

To compute a hypothesis test for the difference between two population proportions, press  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Calculator,  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Statistics,  $\left[\frac{\square}{\square}\right]$  for Stat Tests, and  $\left[\frac{\square}{\square}\right]$  for 2-Prop z-test. Complete the dialog box. Be sure to enter each number of successes as a whole number.

**TI-89**

To calculate a confidence interval for the difference between two population proportions,

- ▶ Select **6:2-PropZInt** from the **STAT Ints** menu.
- ▶ Enter the observed counts and the sample sizes for both samples.
- ▶ Specify a confidence level.
- ▶ Calculate the interval.

To do the mechanics of a hypothesis test for equality of population proportions,

- ▶ Select **6:2-PropZTest** from the **STAT Tests** menu.
- ▶ Enter the observed counts and sample sizes.

- ▶ Indicate what kind of test you want: one-tail upper tail, lower tail, or two-tail.
- ▶ Specify whether results should simply be calculated or displayed with the area corresponding to the P-value of the test shaded.

**COMMENTS**

*Beware:* When you enter the value of  $x$ , you need the *count*, not the percentage. The count must be a whole number. If the number of successes is given as a percent, you must first multiply  $np$  and round the result to obtain  $x$ .

**Chapter 23. Inferences About Means****DATA DESK**

Select variables.  
From the **Calc** menu, choose **Estimate** for confidence intervals or **Test** for hypothesis tests. Select the interval or test

from the drop-down menu and make other choices in the dialog.

**EXCEL**

Specify formulas. Find  $t^*$  with the TINV(alpha, df) function.

**COMMENTS**

Not really automatic. There's no easy way to find P-values in Excel.

**JMP**

From the **Analyze** menu, select **Distribution**. For a confidence interval, scroll down to the “Moments” section to find the interval limits. For a hypothesis test, click the red triangle next to the variable's name and choose **Test Mean** from the menu. Then fill in the resulting dialog.

**COMMENTS**

“Moment” is a fancy statistical term for means, standard deviations, and other related statistics.

**MINITAB**

From the **Stat** menu, choose the **Basic Statistics** submenu. From that menu, choose **1-sample t . . .** Then fill in the dialog.

**COMMENTS**

The dialog offers a clear choice between confidence interval and test.

SPSS	COMMENTS
<p>From the <b>Analyze</b> menu, choose the <b>Compare Means</b> submenu. From that, choose the <b>One-Sample t-test</b> command.</p>	<p>The commands suggest neither a single mean nor an interval. But the results provide both a test and an interval.</p>
TI-NSPIRE	
<p>To compute a confidence interval for a population mean, press <math>\left[\frac{2}{\square}\right]</math>, <math>\left[\frac{1}{\square}\right]</math> for Calculator, <math>\left[\frac{\text{MENU}}{\square}\right]</math>, <math>\left[\frac{6}{\square}\right]</math> for Statistics, <math>\left[\frac{6}{\square}\right]</math> for Confidence Intervals, and <math>\left[\frac{2}{\square}\right]</math> for <i>t</i>-interval. Select between Data and Stats, <math>\left[\frac{\text{TAB}}{\square}\right]</math> to OK, and press <math>\left[\frac{\text{MENU}}{\square}\right]</math>. Complete the dialog box. Be sure to enter the number of successes, <i>x</i>, as a whole number, and the <i>C</i> level as a decimal, such as .99.</p>	<p>To compute a hypothesis test for a population mean, press <math>\left[\frac{2}{\square}\right]</math>, <math>\left[\frac{1}{\square}\right]</math> for Calculator, <math>\left[\frac{\text{MENU}}{\square}\right]</math>, <math>\left[\frac{6}{\square}\right]</math> for Statistics, <math>\left[\frac{7}{\square}\right]</math> for Stat Tests, and <math>\left[\frac{2}{\square}\right]</math> for <i>t</i>-test. Select between Data and Stats, <math>\left[\frac{\text{TAB}}{\square}\right]</math> to OK, and <math>\left[\frac{\text{MENU}}{\square}\right]</math>. Complete the dialog box.</p>
TI-89	
<p>Finding a confidence interval: In the <b>STAT Ints</b> menu, choose <b>2:TInterval</b>. Specify whether you are using data stored in a list or whether you will enter the mean, standard deviation, and sample size. You must also specify the desired level of confidence. Testing a hypothesis: In the <b>STAT Tests</b> menu, choose <b>2:T-Test</b>. You must specify whether you are using data stored in a list or whether you will</p>	<p>enter the mean, standard deviation, and size of your sample. You must also specify the hypothesized model mean and whether the test is to be two-tail, lower-tail, or upper-tail. Select whether the test is to be simply computed or whether to display the distribution curve and highlight the area corresponding to the <i>P</i>-value of the test.</p>

## Chapter 24. Comparing Means

<p>There are two ways to organize data when we want to compare two independent groups. The data can be in two lists, as in the table at the start of this chapter. Each list can be thought of as a variable. In this method, the variables in the batteries example would be <i>Brand Name</i> and <i>Generic</i>. Graphing calculators usually prefer this form, and some computer programs can use it as well.</p>																											
<p>There's another way to think about the data. What is the response variable for the battery life experiment? It's the <i>Time</i> until the music stopped. But the values of this variable are in both columns, and actually there's an experiment factor here, too—namely, the <i>Brand</i> of the battery. So, we could put the data into two different columns, one with the <i>Times</i> in it and one with the <i>Brand</i>. Then the data would look as shown in the table to the right.</p>	<table border="1"> <thead> <tr> <th>Time</th> <th>Brand</th> </tr> </thead> <tbody> <tr><td>194.0</td><td>Brand name</td></tr> <tr><td>205.5</td><td>Brand name</td></tr> <tr><td>199.2</td><td>Brand name</td></tr> <tr><td>172.4</td><td>Brand name</td></tr> <tr><td>184.0</td><td>Brand name</td></tr> <tr><td>169.5</td><td>Brand name</td></tr> <tr><td>190.7</td><td>Generic</td></tr> <tr><td>203.5</td><td>Generic</td></tr> <tr><td>203.5</td><td>Generic</td></tr> <tr><td>206.5</td><td>Generic</td></tr> <tr><td>222.5</td><td>Generic</td></tr> <tr><td>209.4</td><td>Generic</td></tr> </tbody> </table>	Time	Brand	194.0	Brand name	205.5	Brand name	199.2	Brand name	172.4	Brand name	184.0	Brand name	169.5	Brand name	190.7	Generic	203.5	Generic	203.5	Generic	206.5	Generic	222.5	Generic	209.4	Generic
Time	Brand																										
194.0	Brand name																										
205.5	Brand name																										
199.2	Brand name																										
172.4	Brand name																										
184.0	Brand name																										
169.5	Brand name																										
190.7	Generic																										
203.5	Generic																										
203.5	Generic																										
206.5	Generic																										
222.5	Generic																										
209.4	Generic																										
<p>This way of organizing the data makes sense as well. Now the factor and the response variables are clearly visible. You'll have to see which method your program requires. Some packages even allow you to structure the data either way.</p>																											
<p>The commands to do inference for two independent groups on common statistics technology are not always found in obvious places. Here are some starting guidelines.</p>																											

DATA DESK	COMMENTS
<p>Select variables. From the <b>Calc</b> menu, choose <b>Estimate</b> for confidence intervals or <b>Test</b> for hypothesis tests. Select the interval or test from the drop-down menu and make other choices in the dialog.</p>	<p>Data Desk expects the two groups to be in separate variables.</p>
EXCEL	COMMENTS
<p>From the Data Tab, Analysis Group, choose <b>Data Analysis</b>. Alternatively (if the Data Analysis Tool Pack is not installed), in the Formulas Tab, choose More functions &gt; Statistical &gt; TTEST, and specify Type=3 in the resulting dialog. Fill in the cell ranges for the two groups, the hypothesized difference, and the alpha level.</p>	<p>Excel expects the two groups to be in separate cell ranges. Notice that, contrary to Excel's wording, we do not need to assume that the variances are <i>not</i> equal; we simply choose not to assume that they <i>are</i> equal.</p>

**JMP**

From the **Analyze** menu, select **Fit y by x**. Select variables: a **Y, Response** variable that holds the data and an **X, Factor** variable that holds the group names. JMP will make a dotplot. Click the **red triangle** in the dotplot title, and choose **Unequal variances**. The  $t$ -test is at the bottom of the resulting table. Find the P-value from the Prob>F section of the table (they are the same).

**COMMENTS**

JMP expects data in one variable and category names in the other. Don't be misled: There is no need for the variances to be unequal to use two-sample  $t$  methods.

**MINITAB**

From the **Stat** menu, choose the **Basic Statistics** submenu. From that menu, choose **2-sample t....** Then fill in the dialog.

**COMMENTS**

The dialog offers a choice of data in two variables, or data in one variable and category names in the other.

**SPSS**

From the **Analyze** menu, choose the **Compare Means** submenu. From that, choose the **Independent-Samples t-test** command. Specify the data variable and "group variable." Then type in the labels used in the group variable. SPSS offers both the two-sample and pooled- $t$  results in the same table.

**COMMENTS**

SPSS expects the data in one variable and group names in the other. If there are more than two group names in the group variable, only the two that are named in the dialog box will be compared.

**TI-NSPIRE**

To compute a confidence interval for the difference between two population means, press  $\left[\frac{\text{2nd}}{\text{CALC}}$ ,  $\left[\frac{\text{1}}{\text{CI}}$  for Calculator,  $\left[\frac{\text{MEMO}}{\text{STATS}}$ ,  $\left[\frac{\text{6}}{\text{2-SAMP}}$  for Statistics,  $\left[\frac{\text{6}}{\text{CI}}$  for Confidence Intervals, and  $\left[\frac{\text{4}}{\text{2-SAMP}}$  for 2-Sample  $t$ -interval. Select between Data and Stats,  $\left[\frac{\text{tab}}{\text{OK}}$  to OK, and  $\left[\frac{\text{2nd}}{\text{QUIT}}$ . Complete the dialog box. Be sure to enter the C level as a decimal, such as .99.

To compute a hypothesis test for the difference between two population means, press  $\left[\frac{\text{2nd}}{\text{CALC}}$ ,  $\left[\frac{\text{1}}{\text{CI}}$  for Calculator,  $\left[\frac{\text{MEMO}}{\text{STATS}}$ ,  $\left[\frac{\text{6}}{\text{2-SAMP}}$  for Statistics,  $\left[\frac{\text{7}}{\text{STATS}}$  for Stat Tests, and  $\left[\frac{\text{4}}{\text{2-SAMP}}$  for 2-Sample  $t$ -test. Select between Data and Stats,  $\left[\frac{\text{tab}}{\text{OK}}$  to OK, and  $\left[\frac{\text{2nd}}{\text{QUIT}}$ . Complete the dialog box.

**TI-89**

For a confidence interval:  
In the **STAT Ints** menu, choose **4:2-SampTInt**. You must specify if you are using data stored in two lists or if you will enter the means, standard deviations, and sizes of both samples. You must also indicate whether to pool the variances (when in doubt, say no) and specify the desired level of confidence.

To test a hypothesis:  
In the **STAT TESTS** menu, choose **4:2-SampTTest**. You must specify if you are using data stored in two lists or if you will enter the means, standard deviations, and sizes of both samples. You must also indicate whether to pool the variances (when in doubt, say no) and specify whether the test is to be two-tail, lower-tail, or upper-tail.

**Chapter 25. Paired Samples and Blocks****DATA DESK**

Select variables.  
From the **Calc** menu, choose **Estimate** for confidence intervals or **Test** for hypothesis tests. Select the interval or test from the drop-down menu, and make other choices in the dialog.

**COMMENTS**

Data Desk expects the two groups to be in separate variables and in the same "Relation"—that is, about the same cases.

**EXCEL**

In Excel 2003 and earlier, select **Data Analysis** from the **Tools** menu.  
In Excel 2007, select **Data Analysis** from the **Analysis** Group on the **Data** Tab.  
From the **Data Analysis** menu, choose **t-test: paired two-sample for Means**. Fill in the cell ranges for the two groups, the hypothesized difference, and the alpha level.

**COMMENTS**

Excel expects the two groups to be in separate cell ranges.  
**Warning:** Do not compute this test in Excel without checking for missing values. If there are any missing values (empty cells), Excel will usually give a wrong answer. Excel compacts each list, pushing values up to cover the missing cells, and then checks only that it has the same number of values in each list. The result is mismatched pairs and an entirely wrong analysis.

**JMP**

From the **Analyze** menu, select **Matched Pairs**. Specify the columns holding the two groups in the **Y Paired Response** Dialog. Click **OK**.

<p><b>MINITAB</b></p> <p>From the <b>Stat</b> menu, choose the <b>Basic Statistics</b> submenu. From that menu, choose <b>Paired t...</b> Then fill in the dialog.</p>	<p><b>COMMENTS</b></p> <p>Minitab takes “First sample” minus “Second sample.”</p>
<p><b>SPSS</b></p> <p>From the <b>Analyze</b> menu, choose the <b>Compare Means</b> submenu. From that, choose the <b>Paired-Samples t-test</b> command. Select pairs of variables to compare, and click the arrow to add them to the selection box.</p>	<p><b>COMMENTS</b></p> <p>You can compare several pairs of variables at once. Options include the choice to exclude cases missing in any pair from all tests.</p>
<p><b>TI-NSPIRE</b></p> <p>For inference on a matched pair design, compute a third list of differences such as <math>diff = time2 - time1</math>. Then construct the</p>	<p>confidence interval or conduct the hypothesis test in the same way as 1-sample procedures, using the list of differences.</p>
<p><b>TI-89</b></p> <p>If the data are stored in two lists, say, list1 and list2, create a list of the differences: Move the cursor to the name of an empty list, and then use VAR-LINK to enter the command list1-list2. Press <b>[ENTER]</b> to perform the subtraction.</p>	<p>Since inference for paired differences uses one-sample <math>t</math>-procedures, select <b>2:T-Test</b> or <b>2:TInterval</b> from the <b>STAT Tests</b> or <b>Ints</b> menu. Specify as your data the list of differences you just created, and apply the procedure.</p>

## Chapter 26. Comparing Counts

<p><b>DATA DESK</b></p> <p>Select variables. From the <b>Calc</b> menu, choose <b>Contingency Table</b>. From the table’s HyperView menu, choose <b>Table Options</b>. (Or Choose <b>Calc &gt; Calculation Options &gt; Table Options</b>.) In the dialog, check the boxes for <b>Chi Square</b> and for <b>Standardized Residuals</b>. Data Desk will display the chi-square and its P-value below the table, and the standardized residuals within the table.</p>	<p><b>COMMENTS</b></p> <p>Data Desk automatically treats variables selected for this command as categorical variables even if their elements are numerals. The <b>Compute Counts</b> command in the table’s HyperView menu will make variables that hold the table contents (as selected in the Table Options dialog), including the standardized residuals.</p>
<p><b>EXCEL</b></p> <p>Excel offers the function <b>CHITEST(actual_range, expected_range)</b>, which computes a chi-square value for homogeneity. Both ranges are of the form UpperleftCell:LowerRightCell, specifying two rectangular tables that must hold counts (although Excel will not check for integer values). The two tables must be of the same size and shape.</p>	<p><b>COMMENTS</b></p> <p>Excel’s documentation claims this is a test for independence and labels the input ranges accordingly, but Excel offers no way to find expected counts, so the function is not particularly useful for testing independence. You can use this function only if you already know both tables of counts or are willing to program additional calculations.</p>
<p><b>JMP</b></p> <p>From the <b>Analyze</b> menu, select <b>Fit Y by X</b>. Select variables: a Y, Response variable that holds responses for one variable, and an X, Factor variable that holds responses for the other. Both selected variables must be Nominal or Ordinal. JMP will make a plot and a contingency table. Below the contingency table, <b>JMP</b> offers a <b>Tests</b> panel. In that panel, the Chi Square for independence is called a <b>Pearson ChiSquare</b>. The table also offers the P-value. Click on the Contingency Table title bar to drop down a menu that offers to include a <b>Deviation</b> and <b>Cell Chi square</b> in each cell of the table.</p>	<p><b>COMMENTS</b></p> <p>JMP will choose a chi-square analysis for a <b>Fit Y by X</b> if both variables are nominal or ordinal (marked with an N or O), but not otherwise. Be sure the variables have the right type. Deviations are the observed—expected differences in counts. Cell chi-squares are the squares of the standardized residuals. Refer to the deviations for the sign of the difference. Look under <b>Distributions</b> in the <b>Analyze</b> menu to find a chi-square test for goodness-of-fit.</p>

**MINITAB**

From the **Stat** menu, choose the **Tables** submenu. From that menu, choose **Chi Square Test . . .** In the dialog, identify the columns that make up the table. Minitab will display the table and print the chi-square value and its P-value.

**COMMENTS**

Alternatively, select the **Cross Tabulation . . .** command to see more options for the table, including expected counts and standardized residuals.

**SPSS**

From the **Analyze** menu, choose the **Descriptive Statistics** submenu. From that submenu, choose **Crosstabs . . .** In the Crosstabs dialog, assign the row and column variables from the variable list. Both variables must be categorical. Click the **Cells** button to specify that standardized residuals should be displayed. Click the **Statistics** button to specify a chi-square test.

**COMMENTS**

SPSS offers only variables that it knows to be categorical in the variable list for the Crosstabs dialog. If the variables you want are missing, check that they have the right type.

**TI-NSPIRE**

To conduct a  $\chi^2$  goodness of fit test, enter the observed and the expected values into two named lists. Then press  $\left[\frac{\text{2nd}}{\text{CALC}}\right]$ ,  $\left[\frac{\text{1}}{\text{TESTS}}\right]$  for Calculator,  $\left[\frac{\text{MENU}}{\text{STATS}}\right]$ ,  $\left[\frac{\text{6}}{\text{STAT TESTS}}\right]$  for Statistics,  $\left[\frac{\text{7}}{\text{STAT TESTS}}\right]$  for Stat Tests, and  $\left[\frac{\text{7}}{\text{STAT TESTS}}\right]$  for  $\chi^2$  GOF. Complete the dialog box.

To conduct a  $\chi^2$  test of independence or homogeneity, first enter the data into a matrix. Press  $\left[\frac{\text{ctrl}}{\text{MTRX}}\right]$   $\left[\frac{\text{MTRX}}{\text{W}}\right]$  and select the matrix icon.

Enter the dimensions and  $\left[\frac{\text{tab}}{\text{OK}}\right]$  to OK, and  $\left[\frac{\text{ctrl}}{\text{MTRX}}\right]$ . Then type the data into the matrix. Then press  $\left[\frac{\text{ctrl}}{\text{MTRX}}\right]$  to exit the matrix, press  $\left[\frac{\text{ctrl}}{\text{MTRX}}\right]$  and a matrix name such as *ma* to store the matrix. To complete the test, press  $\left[\frac{\text{CALC}}{\text{1}}\right]$  for Calculator,  $\left[\frac{\text{MENU}}{\text{6}}\right]$  for Statistics,  $\left[\frac{\text{7}}{\text{7}}\right]$  for Stat Tests, and  $\left[\frac{\text{8}}{\text{8}}\right]$  for  $\chi^2$  2-way Test. Complete the dialog box.

**TI-89**

To test goodness-of-fit, enter the observed counts in a list and the expected counts in another list. Expected counts can be entered as  $n \cdot p$ , and the calculator will compute them for you. From the **STAT TESTS** menu, select **7:Chi2 GOF**. Enter the list names using VAR-LINK and the degrees of freedom,  $k - 1$ , where  $k$  is the number of categories. Select whether to simply calculate or display the result with the area corresponding to the P-value highlighted.

To test a hypothesis of homogeneity or independence, you need to enter the data as a matrix. From the home screen, press  $\left[\frac{\text{APPS}}{\text{6:Data/Matrix Editor}}\right]$  and select **6:Data/Matrix Editor**, then select **3:New**. Specify type as Matrix and name the matrix in the **Variable** box. Specify the number of rows and columns. Type the entries, pressing  $\left[\frac{\text{ENTER}}{\text{ENTER}}\right]$  after each. Press  $\left[\frac{\text{2nd}}{\text{2nd}}\right]$   $\left[\frac{\text{ESC}}{\text{ESC}}\right]$  to leave the editor. To do the test, choose **8:Chi2 2-way** from the **STAT TESTS** menu.

**Chapter 27. Inferences for Regression****DATA DESK**

- ▶ Select Y- and X-variables.
- ▶ From the **Calc** menu, choose **Regression**.
- ▶ Data Desk displays the regression table.
- ▶ Select plots of residuals from the Regression table's HyperView menu.

**COMMENTS**

You can change the regression by dragging the icon of another variable over either the Y- or X-variable name in the table and dropping it there. The regression will recompute automatically.

**EXCEL**

- ▶ In Excel 2003 and earlier, select Data Analysis from the **Tools** menu. In Excel 2007, select Data Analysis from the **Analysis Group** on the Data Tab.
- ▶ Select Regression from the **Analysis Tools** list.
- ▶ Click the **OK** button.
- ▶ Enter the data range holding the Y-variable in the box labeled "Y-range".
- ▶ Enter the range of cells holding the X-variable in the box labeled "X-range."
- ▶ Select the **New Worksheet Ply** option.
- ▶ Select **Residuals** options. Click the **OK** button.

**COMMENTS**

The Y and X ranges do not need to be in the same rows of the spreadsheet, although they must cover the same number of cells. But it is a good idea to arrange your data in parallel columns as in a data table.

Although the dialog offers a Normal probability plot of the residuals, the data analysis add-in does not make a correct probability plot, so don't use this option.

**JMP**

- ▶ From the **Analyze** menu, select **Fit Y by X**.
- ▶ Select variables: a Y, Response variable, and an X, Factor variable. Both must be continuous (quantitative).
- ▶ JMP makes a scatterplot.
- ▶ Click on the red triangle beside the heading labeled **Bivariate Fit...** and choose **Fit Line**. JMP draws the least squares regression line on the scatterplot and displays the results of the regression in tables below the plot.
- ▶ The portion of the table labeled “Parameter Estimates” gives the coefficients and their standard errors, *t*-ratios, and P-values.

**COMMENTS**

JMP chooses a regression analysis when both variables are “Continuous.” If you get a different analysis, check the variable types.  
The Parameter table does not include the residual standard deviation  $s_e$ . You can find that as Root Mean Square Error in the Summary of Fit panel of the output.

**MINITAB**

- ▶ Choose **Regression** from the **Stat** menu.
- ▶ Choose **Regression...** from the **Regression** submenu.
- ▶ In the Regression dialog, assign the Y-variable to the Response box and assign the X-variable to the Predictors box.
- ▶ Click the **Graphs** button.
- ▶ In the Regression-Graphs dialog, select **Standardized residuals**, and check **Normal plot of residuals** and **Residuals versus fits**.

- ▶ Click the **OK** button to return to the Regression dialog.
- ▶ Click the **OK** button to compute the regression.

**COMMENTS**

You can also start by choosing a Fitted Line plot from the **Regression** submenu to see the scatterplot first—usually good practice.

**SPSS**

- ▶ Choose **Regression** from the **Analyze** menu.
- ▶ Choose **Linear** from the **Regression** submenu.
- ▶ In the Linear Regression dialog that appears, select the Y-variable and move it to the dependent target. Then move the X-variable to the independent target.
- ▶ Click the **Plots** button.

- ▶ In the Linear Regression Plots dialog, choose to plot the \*SRESIDs against the \*ZPRED values.
- ▶ Click the **Continue** button to return to the Linear Regression dialog.
- ▶ Click the **OK** button to compute the regression.

**TI-NSPIRE**

To compute a confidence interval for a population slope, first enter the data into two named lists. Then press  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Calculator,  $\left[\text{MENU}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Statistics,  $\left[\frac{\square}{\square}\right]$  for Confidence Intervals, and  $\left[\frac{\square}{\square}\right]$  for Linear Reg *t*-intervals. Select slope,  $\left[\text{tab}\right]$  to OK, and  $\left[\frac{\square}{\square}\right]$ . Complete the dialog box. Be sure to enter the C level as a decimal, such as .99.

To compute a hypothesis test for a population slope, first enter the data into two named lists. Then press  $\left[\frac{\square}{\square}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Calculator,  $\left[\text{MENU}\right]$ ,  $\left[\frac{\square}{\square}\right]$  for Statistics,  $\left[\frac{\square}{\square}\right]$  for Stat Tests, and  $\left[\text{A}\right]$  for Linear Reg *t*-test. Complete the dialog box.

**TI-89**

Under **STAT Tests** choose **A:LinRegTTest**. Specify the two lists where the data are stored and (usually) choose the two-tail option. Select an equation name to store the resulting line. In addition to reporting the calculated value of *t* and the P-value, the calculator will tell you the coefficients of the regression equation (*a* and *b*), the values of  $r^2$  and *r*, the value of *s* used in predic-

tion and confidence intervals, and the standard error of the slope. For 95% prediction and confidence intervals, choose **7:LinRegTint** from the **STAT Ints** menu. Specify the two lists where the data are stored, and select an equation name to store the resulting line. Select for an interval for the slope or for a response. If for a response, enter the *x*-value.

**Chapter 28. Analysis of Variance****DATA DESK**

- ▶ Select the response variable as Y and the factor variable as X.
- ▶ From the **Calc** menu, choose **ANOVA**.
- ▶ Data Desk displays the ANOVA table.
- ▶ Select plots of residuals from the ANOVA table’s HyperView menu.

**COMMENTS**

Data Desk expects data in “stacked” format. You can change the ANOVA by dragging the icon of another variable over either the Y or X variable name in the table and dropping it there. The analysis will recompute automatically.

**EXCEL**

- ▶ In Excel 2003 and earlier, select **Data Analysis** from the Tools menu.
- ▶ In Excel 2007, select **Data Analysis** from the Analysis Group on the Data Tab.
- ▶ Select **Anova Single Factor** from the list of analysis tools.
- ▶ Click the **OK** button.
- ▶ Enter the data range in the box provided.
- ▶ Check the **Labels in First Row** box, if applicable.
- ▶ Enter an alpha level for the F-test in the box provided.
- ▶ Click the **OK** button.

**COMMENTS**

The data range should include two or more columns of data to compare. Unlike all other statistics packages, Excel expects each column of the data to represent a different level of the factor. However, it offers no way to label these levels. The columns need not have the same number of data values, but the selected cells must make up a rectangle large enough to hold the column with the most data values.

**JMP**

- ▶ From the **Analyze** menu select **Fit Y by X**.
- ▶ Select variables: a quantitative Y, Response variable, and a categorical X, Factor variable.
- ▶ JMP opens the **Oneway** window.
- ▶ Click on the red triangle beside the heading, select **Display Options**, and choose **Boxplots**.

- ▶ From the same menu choose the **Means/ANOVA.t-test** command.
- ▶ JMP opens the oneway ANOVA output.

**COMMENTS**

JMP expects data in “stacked” format with one response and one factor variable.

**MINITAB**

- ▶ Choose **ANOVA** from the Stat menu.
- ▶ Choose **One-way...** from the **ANOVA** submenu.
- ▶ In the One-way Anova dialog, assign a quantitative Y variable to the Response box and assign a categorical X variable to the Factor box.
- ▶ Check the **Store Residuals** check box.
- ▶ Click the **Graphs** button.
- ▶ In the ANOVA-Graphs dialog, select **Standardized residuals**, and check **Normal plot of residuals** and **Residuals versus fits**.

- ▶ Click the **OK** button to return to the Regression dialog.
- ▶ Click the **OK** button to compute the regression.

**COMMENTS**

If your data are in unstacked format, with separate columns for each treatment level, choose **One-way (unstacked)** from the **ANOVA** submenu.

**SPSS**

- ▶ Choose **Compare Means** from the **Analyze** menu.
- ▶ Choose **One-way ANOVA** from the **Compare Means** submenu.
- ▶ In the One-Way ANOVA dialog, select the Y-variable and move it to the dependent target. Then move the X-variable to the independent target.
- ▶ Click the **OK** button.

**COMMENTS**

SPSS expects data in stacked format. The **Contrasts** and **Post Hoc** buttons offer ways to test contrasts and perform multiple comparisons. See your SPSS manual for details.

**TI-89**

Under **STAT Tests**, choose **C:ANOVA**

- ▶ Specify the input method (Data or Stats) according to whether you have data entered as one list for each group or summary statistics for each group, and specify the number of groups. Press  $\div$ .
- ▶ If Data, you will then be asked to supply the name of each list.
- ▶ If Stats, you will be asked for the stats for each group. Enter  $n$ ,  $\bar{x}$ , and  $s$  for each group separated by commas and within curly braces ( $\{\}$  and  $\}$ ).
- ▶ Press  $\div$  to perform the calculations.

**COMMENTS**

In addition to the ANOVA table output, the calculator creates three new lists—the means for each group (in the order specified) and *individual* 95% confidence interval upper and lower bounds.

## Chapter 29. Multiple Regression

### DATA DESK

- ▶ Select Y- and X-variable icons.
- ▶ From the **Calc** menu, choose **Regression**.
- ▶ Data Desk displays the regression table.
- ▶ Select plots of residuals from the Regression table's HyperView menu.

### COMMENTS

You can change the regression by dragging the icon of another variable over either the Y- or an X-variable name in the table and dropping it there. You can add a predictor by dragging its icon into that part of the table. The regression will recompute automatically.

### EXCEL

- ▶ In Excel 2003 and earlier, select **Data Analysis** from the **Tools** menu.
- ▶ In Excel 2007, select **Data Analysis** from the **Analysis Group** on the Data Tab.
- ▶ Select **Regression** from the **Analysis Tools** list.
- ▶ Click the **OK** button.
- ▶ Enter the data range holding the Y-variable in the box labeled "Y-range."
- ▶ Enter the range of cells holding the X-variables in the box labeled "X-range."
- ▶ Select the **New Worksheet Ply** option.
- ▶ Select **Residuals** options. Click the **OK** button.

### COMMENTS

The Y and X ranges do not need to be in the same rows of the spreadsheet, although they must cover the same number of cells. But it is a good idea to arrange your data in parallel columns as in a data table. The X-variables must be in adjacent columns. No cells in the data range may hold non-numeric values.

Although the dialog offers a Normal probability plot of the residuals, the data analysis add-in does not make a correct probability plot, so don't use this option.

### JMP

- ▶ From the **Analyze** menu select **Fit Model**.
- ▶ Specify the response, Y. Assign the predictors, X, in the **Construct Model Effects** dialog box.
- ▶ Click on **Run Model**.

### COMMENTS

JMP chooses a regression analysis when the response variable is "Continuous." The predictors can be any combination of quantitative or categorical. If you get a different analysis, check the variable types.

### MINITAB

- ▶ Choose **Regression** from the **Stat** menu.
- ▶ Choose **Regression . . .** from the **Regression** submenu.
- ▶ In the Regression dialog, assign the Y-variable to the Response box and assign the X-variables to the Predictors box.
- ▶ Click the **Graphs** button.

- ▶ In the Regression-Graphs dialog, select **Standardized residuals**, and check **Normal plot of residuals** and **Residuals versus fits**.
- ▶ Click the **OK** button to return to the Regression dialog.
- ▶ Click the **OK** button to compute the regression.

### SPSS

- ▶ Choose **Regression** from the **Analyze** menu.
- ▶ Choose **Linear** from the **Regression** submenu.
- ▶ When the Linear Regression dialog appears, select the Y-variable and move it to the dependent target. Then move the X-variables to the independent target.
- ▶ Click the **Plots** button.

- ▶ In the Linear Regression Plots dialog, choose to plot the \*SRESIDs against the \*ZPRED values.
- ▶ Click the **Continue** button to return to the Linear Regression dialog.
- ▶ Click the **OK** button to compute the regression.

### TI-89

Under **STAT Tests** choose **B:MultReg Tests**

- ▶ Specify the number of predictor variables, and which lists contain the response variable and predictor variables.
- ▶ Press  $\square$  to perform the calculations.

### COMMENTS

- ▶ The first portion of the output gives the  $F$ -statistic and its P-value as well as the values of  $R^2$ ,  $Adj^2R^2$ , the standard deviation of the residuals ( $s$ ), and the Durbin-Watson statistic, which measures correlation among the residuals.

- ▶ The rest of the main output gives the components of the  $F$ -test, as well as values of the coefficients, their standard errors, and associated  $t$ -statistics along with P-values. You can use the right arrow to scroll through these lists (if desired).
- ▶ The calculator creates several new lists that can be used for assessing the model and its conditions: Yhatlist, resid, sresid (standardized residuals), leverage, and cookd, as well as lists of the coefficients, standard errors,  $t$ 's, and P-values.



# Answers

## APPENDIX

# C

Here are the “answers” to the exercises for the chapters and the unit reviews. As we said in Chapter 1, the answers are outlines of the complete solution. Your solution should follow the model of the Step-By-Step examples, where appropriate. You should explain the context, show your reasoning and calculations, and draw conclusions. For some problems, what you decide to include in an argument may differ somewhat from the answers here. But, of course, the numerical part of your answer should match the numbers in the answers shown.

### CHAPTER 2

1. Categorical
3. Quantitative
5. Answers will vary.
7. *Who*—2500 cars  
*What*—Distance from car to bicycle  
*Population*—All cars passing bicyclists
9. *Who*—Coffee drinkers at a Newcastle University coffee station  
*What*—Amount of money contributed  
*Population*—All people in honor system payment situations
11. *Who*—25,892 men aged 30 to 87  
*What*—Fitness level and cause of death  
*Population*—All men
13. *Who*—54 bears  
*Cases*—Each bear is a case.  
*What*—Weight, neck size, length, and sex  
*When*—Not specified  
*Where*—Not specified  
*Why*—To estimate weight from easier-to-measure variables  
*How*—Researchers collected data on 54 bears they were able to catch.  
*Variable*—Weight  
*Type*—Quantitative  
*Units*—Not specified  
*Variable*—Neck size  
*Type*—Quantitative  
*Units*—Not specified  
*Variable*—Length  
*Type*—Quantitative  
*Units*—Not specified  
*Variable*—Sex  
*Type*—Categorical
15. *Who*—Arby’s sandwiches  
*Cases*—Each sandwich is a case.  
*What*—Type of meat, number of calories, and serving size  
*When*—Not specified  
*Where*—Arby’s restaurants  
*Why*—To assess nutritional value of sandwiches  
*How*—Report by Arby’s restaurants  
*Variable*—Type of meat  
*Type*—Categorical  
*Variable*—Number of calories  
*Type*—Quantitative  
*Units*—Calories  
*Variable*—Serving size  
*Type*—Quantitative  
*Units*—Ounces
17. *Who*—882 births  
*Cases*—Each of the 882 births is a case.  
*What*—Mother’s age, length of pregnancy, type of birth, level of prenatal care, birth weight of baby, sex of baby, and baby’s health problems  
*When*—1998–2000  
*Where*—Large city hospital  
*Why*—Researchers were investigating the impact of prenatal care on newborn health.  
*How*—Not specified exactly, but probably from hospital records  
*Variable*—Mother’s age  
*Type*—Quantitative  
*Units*—Not specified; probably years  
*Variable*—Length of pregnancy  
*Type*—Quantitative  
*Units*—Weeks  
*Variable*—Birth weight of baby  
*Type*—Quantitative  
*Units*—Not specified, probably pounds and ounces  
*Variable*—Type of birth  
*Type*—Categorical  
*Variable*—Level of prenatal care  
*Type*—Categorical  
*Variable*—Sex  
*Type*—Categorical  
*Variable*—Baby’s health problems  
*Type*—Categorical
19. *Who*—Experiment subjects  
*Cases*—Each subject is an individual.  
*What*—Treatment (herbal cold remedy or sugar solution) and cold severity  
*When*—Not specified

Where—Not specified  
 Why—To test efficacy of herbal remedy on common cold  
 How—The scientists set up an experiment.  
 Variable—Treatment  
 Type—Categorical  
 Variable—Cold severity rating  
 Type—Quantitative (perhaps ordinal categorical)  
 Units—Scale from 0 to 5  
 Concerns—The severity of a cold seems subjective and difficult to quantify. Scientists may feel pressure to report negative findings of herbal product.

21. Who—Streams  
 Cases—Each stream is a case.  
 What—Name of stream, substrate of the stream, acidity of the water, temperature, BCI  
 When—Not specified  
 Where—Upstate New York  
 Why—To study ecology of streams  
 How—Not specified  
 Variable—Stream name  
 Type—Identifier  
 Variable—Substrate  
 Type—Categorical  
 Variable—Acidity of water  
 Type—Quantitative  
 Units—pH  
 Variable—Temperature  
 Type—Quantitative  
 Units—Degrees Celsius  
 Variable—BCI  
 Type—Quantitative  
 Units—Not specified

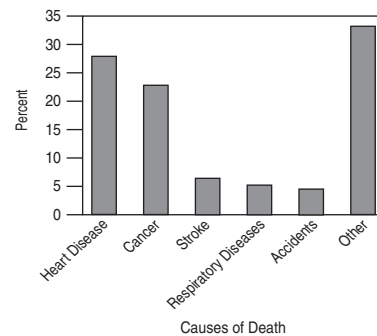
23. Who—41 refrigerator models  
 Cases—Each of the 41 refrigerator models is a case.  
 What—Brand, cost, size, type, estimated annual energy cost, overall rating, and repair history  
 When—2006  
 Where—United States  
 Why—To provide information to the readers of *Consumer Reports*  
 How—Not specified  
 Variable—Brand  
 Type—Categorical  
 Variable—Cost  
 Type—Quantitative  
 Units—Not specified (dollars)  
 Variable—Size  
 Type—Quantitative  
 Units—Cubic feet  
 Variable—Type  
 Type—Categorical  
 Variable—Estimated annual energy cost  
 Type—Quantitative  
 Units—Not specified (dollars)  
 Variable—Overall rating  
 Type—Categorical (ordinal)  
 Variable—Percent requiring repair in last 5 years  
 Type—Quantitative  
 Units—Percent

25. Who—Kentucky Derby races  
 What—Date, winner, margin, jockey, net proceed to winner, duration, track condition  
 When—1875 to 2008  
 Where—Churchill Downs, Louisville, Kentucky  
 Why—Not specified (To see trends in horse racing?)  
 How—Official statistics collected at race  
 Variable—Year  
 Type—Quantitative

Units—Day and year  
 Variable—Winner  
 Type—Identifier  
 Variable—Margin  
 Type—Quantitative  
 Units—Horse lengths  
 Variable—Jockey  
 Type—Categorical  
 Variable—Net proceeds to winner  
 Type—Quantitative  
 Units—Dollars  
 Variable—Duration  
 Type—Quantitative  
 Units—Minutes and seconds  
 Variable—Track condition  
 Type—Categorical

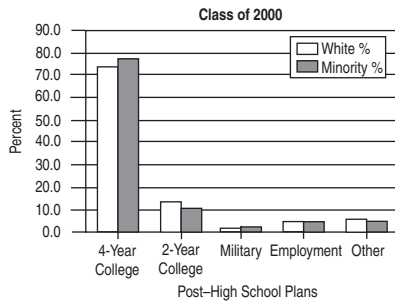
### CHAPTER 3

- Answers will vary.
- Answers will vary.
- a) Yes; each is categorized in a single genre.  
 b) Thriller/Horror
- a) Comedy  
 b) It is easier to tell from the bar chart; slices of the pie chart are too close in size.
- 1755 students applied for admission to the magnet schools program. 53% were accepted, 17% were wait-listed, and the other 30% were turned away.
- a) Yes. We can add because these categories do not overlap. (Each person is assigned only one cause of death.)  
 b)  $100 - (27.2 + 23.1 + 6.3 + 5.1 + 4.7) = 33.6\%$   
 c) Either a bar chart or pie chart with “other” added would be appropriate. A bar chart is shown.



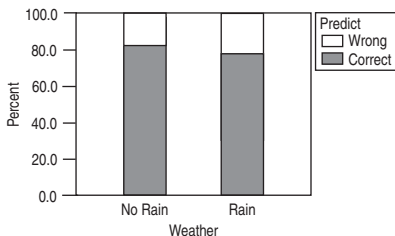
- a) The bar chart shows that grounding and collision are the most frequent causes of oil spills. Very few have unknown causes.  
 b) A pie chart seems appropriate as well.
- There’s no title, the percentages total only 92%, and the three-dimensional display distorts the sizes of the regions.
- In both the South and West, about 58% of the eighth-grade smokers preferred Marlboro. Newport was the next most popular brand, but was far more popular in the South than in the West, where Camel was cited nearly 3 times as often as in the South. Nearly twice as many smokers in the West as in the South indicated that they had no usual brand (12.9% to 6.7%).
- a) The column totals are 100%.  
 b) 31.7%  
 c) 60%  
 d) i. 35.7%; ii. can’t tell; iii. 0%; iv. can’t tell
- a) 82.5%    b) 12.9%    c) 11.1%  
 d) 13.4%    e) 85.7%
- a) 73.9% 4-yr college, 13.4% 2-year college, 1.5% military, 5.2% employment, 6.0% other

- b) 77.2% 4-yr college, 10.5% 2-year college, 1.8% military, 5.3% employment, 5.3% other
- c) Many charts are possible. Here is a side-by-side bar chart.



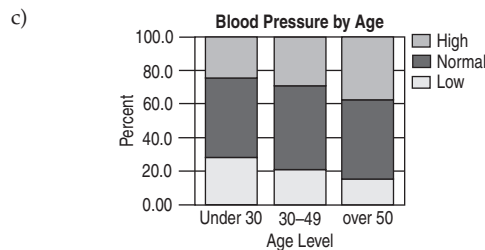
- d) The white and minority students' plans are very similar. The small differences should be interpreted with caution because the total number of minority students is small. There is little evidence of an association between race and plans.

- 25. a) 16.6%      b) 11.8%      c) 37.7%      d) 53.0%
- 27. 1755 students applied for admission to the magnet schools program: 53% were accepted, 17% were wait-listed, and the other 30% were turned away. While the overall acceptance rate was 53%, 93.8% of blacks and Hispanics were accepted, compared to only 37.7% of Asians and 35.5% of whites. Overall, 29.5% of applicants were black or Hispanic, but only 6% of those turned away were. Asians accounted for 16.6% of all applicants, but 25.4% of those turned away. Whites were 54% of the applicants and 68.5% of those who were turned away. It appears that the admissions decisions were not independent of the applicant's ethnicity.
- 29. a) 9.3%      b) 24.7%      c) 80.8%
- d) No, there appears to be no association between weather and ability to forecast weather. On days it rained, his forecast was correct 79.4% of the time. When there was no rain, his forecast was correct 81.0% of the time.

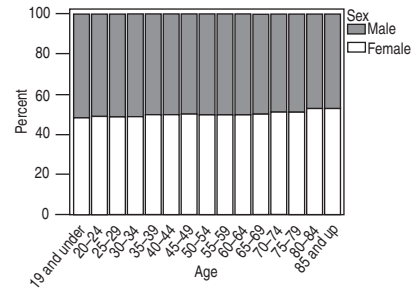


- 31. a) Low 20.0%, Normal 48.9%, High 31.0%
- b)

Blood Pressure	Under 30	30-49	Over 50
	Low	27.6%	20.7%
Normal	49.0%	50.8%	47.2%
High	23.5%	28.5%	37.1%



- d) As age increases, the percent of adults with high blood pressure increases. By contrast, the percent of adults with low blood pressure decreases.
- e) No, but it gives an indication that it might. There might be additional reasons that explain the differences in blood pressures.
- 33. No, there's no evidence that Prozac is effective. The relapse rates were nearly identical: 28.6% among the people treated with Prozac, compared to 27.3% among those who took the placebo.
- 35. a) 4.7%      b) 50.0%
- c) There are about 50% of each sex in each age group, but it ranges from 48.8% female in the youngest group to 54.6% in the oldest. As the age increases, there is a slight increase in the percentage of female drivers.



- d) There is a slight association. As the age increases, there is a small increase in the percentage of female drivers.
- 37. a) 160 of 1300, or 12.3%
- b) Yes. Major surgery: 15.3% vs. minor surgery: 6.7%
- c) Large hospital: 13%; small hospital: 10%
- d) Large hospital: Major 15% vs. minor 5%  
Small hospital: Major 20% vs. minor 8%
- e) No. Smaller hospitals have a higher rate for both kinds of surgery, even though it's lower "overall."
- f) The small hospital has a larger percentage of minor surgeries (83.3%) than the large hospital (20%). Minor surgeries have a lower delay rate, so the small hospital looks better "overall."
- 39. a) 42.6%
- b) A higher percentage of males than females were admitted:  
Males: 47.2% to females: 30.9%
- c) Program 1: Males 61.9%, females 82.4%  
Program 2: Males 62.9%, females 68.0%  
Program 3: Males 33.7%, females 35.2%  
Program 4: Males 5.9%, females 7.0%
- d) The comparisons in c) show that males have a lower admittance rate in every program, even though the overall rate shows males with a higher rate of admittance. This is an example of Simpson's paradox.

**CHAPTER 4**

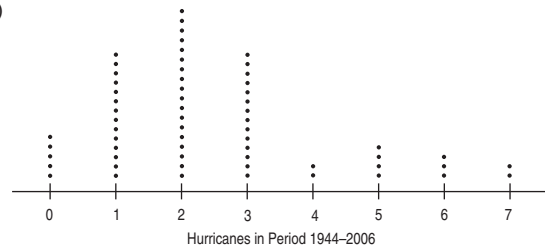
- 1. Answers will vary.
- 3. Answers will vary.
- 5. a) Unimodal (near 0) and skewed to the right. Many seniors will have 0 or 1 speeding tickets. Some may have several, and a few may have more than that.
- b) Probably unimodal and slightly skewed to the right. It is easier to score 15 strokes over the mean than 15 strokes under the mean.
- c) Probably unimodal and symmetric. Weights may be equally likely to be over or under the average.
- d) Probably bimodal. Men's and women's distributions may have different modes. It may also be skewed to the right, since it is possible to have very long hair, but hair length can't be negative.

7. a) Bimodal. Looks like two groups. Modes are near 6% and 46%. No real outliers.  
 b) Looks like two groups of cereals, a low-sugar and a high-sugar group.
9. a) 78%  
 b) Skewed to the right with at least one high outlier. Most of the vineyards are less than 90 acres with a few high ones. The mode is between 0 and 30 acres.
11. a) Because the distribution is skewed to the right, we expect the mean to be larger.  
 b) Bimodal and skewed to the right. Center mode near 8 days. Another mode at 1 day (may represent patients who didn't survive). Most of the patients stay between 1 and 15 days. There are some extremely high values above 25 days.  
 c) The median and IQR, because the distribution is strongly skewed.
13. a) 45 points    b) 37 points and 54 (or 55) points  
 c) In the Super Bowl teams typically score a total of about 45 points, with half the games totaling between 37 and 55 points. In only one fourth of the games have the teams scored fewer than 27 points, and they once totaled 75.
15. a) The standard deviation will be larger for set 2, since the values are more spread out.  $SD(\text{set 1}) = 2.2$ ,  $SD(\text{set 2}) = 3.2$ .  
 b) The standard deviation will be larger for set 2, since 11 and 19 are farther from 15 than are 14 and 16. Other numbers are the same.  $SD(\text{set 1}) = 3.6$ ,  $SD(\text{set 2}) = 4.5$ .  
 c) The standard deviation will be the same for both sets, since the values in the second data set are just the values in the first data set + 80. The spread has not changed.  $SD(\text{set 1}) = 4.2$ ,  $SD(\text{set 2}) = 4.2$ .
17. The mean and standard deviation because the distribution is unimodal and symmetric.
19. a) The mean is closest to \$2.60 because that's the balancing point of the histogram.  
 b) The standard deviation is closest to \$0.15 since that's a typical distance from the mean. There are no prices as far as \$0.50 or \$1.00 from the mean.
21. a) About 100 minutes  
 b) Yes, only 4 of these movies run that long.  
 c) The mean would be higher. The distribution is skewed high.
23. a) i. The middle 50% of movies ran between 97 and 119 minutes.  
 ii. On average, movie lengths varied from the mean run time by 19.6 minutes.  
 b) We should be cautious in using the standard deviation because the distribution of run times is skewed to the right.
25. a) The median will probably be unaffected. The mean will be larger.  
 b) The range and standard deviation will increase; the IQR will be unaffected.
27. The publication is using the median; the watchdog group is using the mean, pulled higher by the several very expensive movies in the long right tail.
29. a) Mean \$525, median \$450  
 b) 2 employees earn more than the mean.  
 c) The median because of the outlier.  
 d) The IQR will be least sensitive to the outlier of \$1200, so it would be the best to report.
31. a)

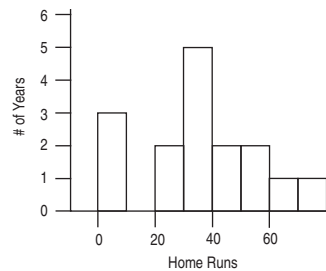
Stem	Leaf
25	
25	
24	56
24	
23	68
23	23
22	677789
22	1234

22|1 = \$2.21/gallon

- b) The distribution of gas prices is unimodal and skewed to the right (upward), centered around \$2.27, with most stations charging between \$2.26 and \$2.33 per gallon. The lowest and highest prices were \$2.21 and \$2.46.  
 c) There are two high prices separated from the other gas stations by a gap.
33. a) Since these data are strongly skewed to the right, the median and IQR are the best statistics to report.  
 b) The mean will be larger than the median because the data are skewed to the right.  
 c) The median is 4 million. The IQR is 4.5 million ( $Q3 = 6$  million,  $Q1 = 1.5$  million).  
 d) The distribution of populations of the states and Washington, DC, is unimodal and skewed to the right. The median population is 4 million. One state is an outlier, with a population of 34 million.
35. Skewed to the right, median at 36. Three low outliers, then a gap from 9 to 22.
37. a)



- b) Slightly skewed to the right. Unimodal, mode near 2. Possibly a second mode near 5. No outliers.
39. a) This is not a histogram. The horizontal axis should split the number of home runs hit in each year into bins. The vertical axis should show the number of years in each bin.  
 b)

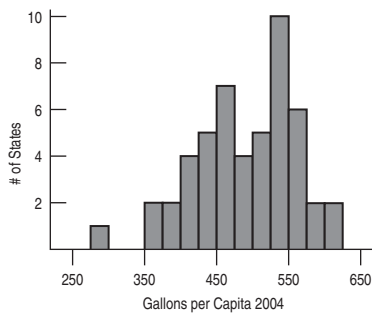


41. Skewed to the right, possibly bimodal with one fairly symmetric group near 4.4, another at 5.6. Two outliers in middle seem not to belong to either group.

Stem	Leaf
57	8
56	27
55	1
54	
53	
52	9
51	
50	8
49	
48	2
47	3
46	034
45	267
44	015
43	0199
42	669
41	22

41|2 = 4.12 pH

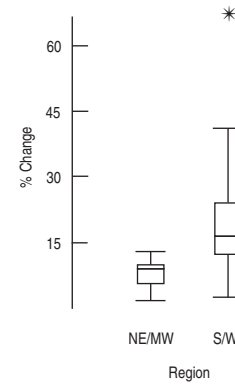
43. Histogram bins are too wide to be useful.
45. Neither appropriate nor useful. Zip codes are categorical data, not quantitative. But they do contain *some* information. The leading digit gives a rough East-to-West placement in the United States. So, we see that they have almost no customers in the Northeast, but a bar chart by leading digit would be more appropriate.
- 47) a) Median 239, IQR 9, Mean 237.6, SD 5.7  
 b) Because it's skewed to the left, probably better to report Median and IQR.  
 c) Skewed to the left; may be bimodal. The center is around 239. The middle 50% of states scored between 233 and 242. Alabama, Mississippi, and New Mexico scores were much lower than other states' scores.
49. In the year 2004, per capita gasoline use by state in the United States averaged around 500 gallons per person (mean 488.8, median 500.5). States varied in per capita consumption, with a standard deviation of 68.7 gallons. The only outlier is New York. The IQR of 96.9 gallons shows that 50% of the states had per capita consumption of between 447.5 and 544.4 gallons. The data appear to be bimodal, so the median and IQR are better choices of summary statistics.



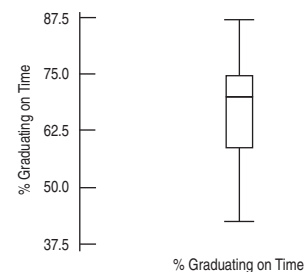
**CHAPTER 5**

1. Answers will vary.
3. Answers will vary.
5. a) Prices appear to be both higher on average and more variable in Baltimore than in the other three cities. Prices in Chicago may be slightly higher than in Dallas and Denver, but the difference is very small.  
 b) There are outliers on the low end in Baltimore and Chicago and one high outlier in Dallas, but these do not affect the overall conclusions reached in part a).  
 7. a) Essentially symmetric, very slightly skewed to the right with two high outliers at 36 and 48. Most victims are between the ages of 16 and 24.  
 b) The slight increase between ages 22 and 24 is apparent in the histogram but not in the boxplot. It may be a second mode.  
 c) The median would be the most appropriate measure of center because of the slight skew and the extreme outliers.  
 d) The IQR would be the most appropriate measure of spread because of the slight skew and the extreme outliers.  
 9. a) About 59%    b) Bimodal  
 c) Some cereals are very sugary; others are healthier low-sugar brands.  
 d) Yes  
 e) Although the ranges appear to be comparable for both groups (about 28%), the IQR is larger for the adult cereals, indicating that there's more variability in the sugar content of the middle 50% of adult cereals.

11. a)



- b) Growth rates in NE/MW states are tightly clustered near 5%. S/W states are more variable, and bimodal with modes near 14 and 22. The S/W states have an outlier as well. Around all the modes, the distributions are fairly symmetric.
13. a) They should be put on the same scale, from 0 to 20 days.  
 b) Lengths of men's stays appear to vary more than for women. Men have a mode at 1 day and then taper off from there. Women have a mode near 5 days, with a sharp drop afterward.  
 c) A possible reason is childbirth.
15. a) Both girls have a median score of about 17 points per game, but Scyrine is much more consistent. Her IQR is about 2 points, while Alexandra's is over 10.  
 b) If the coach wants a consistent performer, she should take Scyrine. She'll almost certainly deliver somewhere between 15 and 20 points. But if she wants to take a chance and needs a "big game," she should take Alexandra. Alex scores over 24 points about a quarter of the time. (On the other hand, she scores under 11 points as often.)
17. Women appear to marry about 3 years younger than men, but the two distributions are very similar in shape and spread.
19. (Note: Numerical details may vary.) In general, fuel economy is higher in cars than in either SUVs or vans. There are numerous outliers on both ends for cars and a few high outliers for SUVs. The top 50% of cars gets higher fuel economy than 75% of SUVs and nearly all vans. On average, SUVs and vans get about the same fuel economy, although the distribution for vans shows less spread. The range for vans is about 40 mpg, while for SUVs it is nearly 30 mpg.
21. The class A is 1, class B is 2, and class C is 3.
23. a) Probably slightly left skewed. The mean is slightly below the median, and the 25th percentile is farther from the median than the 75th percentile.  
 b) No, all data are within the fences.  
 c)

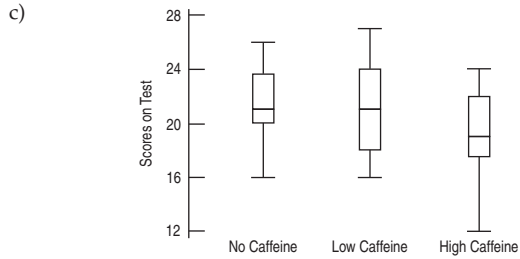


- d) The 48 universities graduate, on average, about 68% of freshmen "on time," with percents ranging from 43% to 87%. The middle 50% of these universities graduate between 59% and 75% of their freshmen in 4 years.
25. a) *Who:* Student volunteers  
*What:* Memory test  
*Where, when:* Not specified

*How:* Students took memory test 2 hours after drinking caffeine-free, half-dose caffeine, or high-caffeine soda.

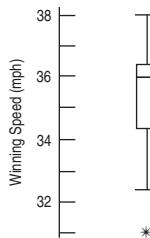
*Why:* To see if caffeine makes you more alert and aids memory retention.

b) Drink: categorical; Test score: quantitative.



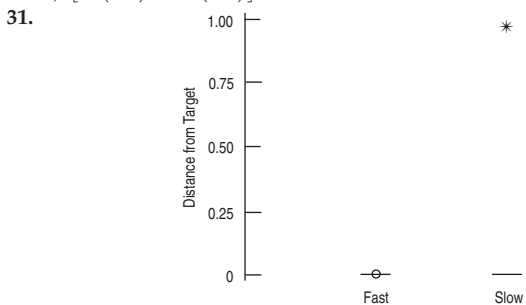
d) The participants scored about the same with no caffeine and low caffeine. The medians for both were 21 points, with slightly more variation for the low-caffeine group. The high-caffeine group generally scored lower than the other two groups on all measures of the 5-number summary: min, lower quartile, median, upper quartile, and max.

27. a) About 36 mph  
 b)  $Q_1$  about 35 mph and  $Q_3$  about 37 mph  
 c) The range appears to be about 7 mph, from about 31 to 38 mph. The IQR is about 2 mph.  
 d) We can't know exactly, but the boxplot may look something like this:

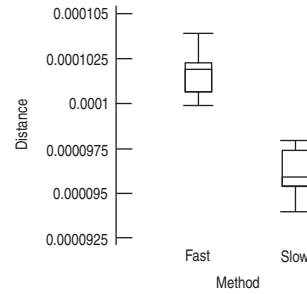


e) The median winning speed has been about 36 mph, with a max of about 38 and a min of about 31 mph. Half have run between about 35 and 37 mph, for an IQR of 2 mph.

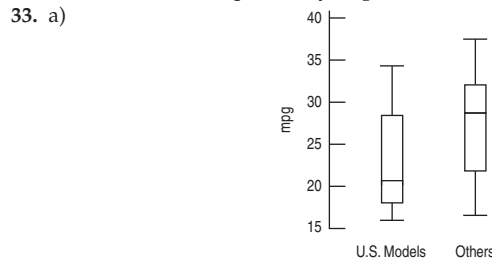
29. a) Boys b) Boys c) Girls  
 d) The boys appeared to have more skew, as their scores were less symmetric between quartiles. The girls' quartiles are the same distance from the median, although the left tail stretches a bit farther to the left.  
 e) Girls. Their median and upper quartiles are larger. The lower quartile is slightly lower, but close.  
 f)  $[14(4.2) + 11(4.6)]/25 = 4.38$



There appears to be an outlier! This point should be investigated. We'll proceed by redoing the plots with the outlier omitted:



It appears that slow speed provides much greater accuracy. But the outlier should be investigated. It is possible that slow speed can induce an infrequent very large distance.

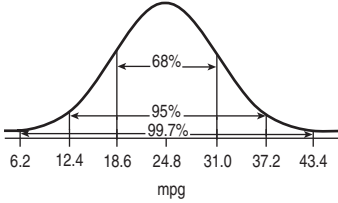


33. a) b) Mileage for U.S. models is typically lower, although the variability is about the same as for cars made elsewhere. The median for U.S. models is around 21 mpg, compared to 28 for the others. Half of U.S. models fall below the first quartile of others. (Other answers possible.)  
 35. a) Day 16 (but any estimate near 20 is okay).  
 b) Day 65 (but anything around 60 is okay).  
 c) Around day 50  
 37. a) Most of the data are found in the far left of this histogram. The distribution is very skewed to the right.  
 b) Re-expressing the data by, for example, logs or square roots might help make the distribution more nearly symmetric.  
 39. a) The logarithm makes the histogram more symmetric. It is easy to see that the center is around 3.5 in log assets.  
 b) That has a value of around 2,500 million dollars.  
 c) That has a value of around 1,000 million dollars.  
 41. a) Fusion time and group.  
 b) Fusion time is quantitative (units = seconds). Group is categorical.  
 c) Both distributions are skewed to the right with high outliers. The boxplot indicates that visual information may reduce fusion time. The median for the Verbal/Visual group seems to be about the same as the lower quartile of the No/Verbal group.

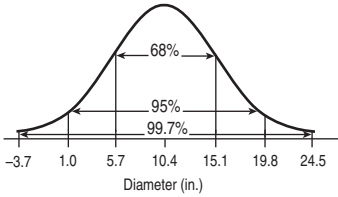
## CHAPTER 6

1. a) 72 oz., 40 oz. b) 4.5 lb, 2.5 lb  
 3. a) Skewed to the right; mean is higher than median.  
 b) \$350 and \$950.  
 c) Minimum \$350. Mean \$750. Median \$550. Range \$1200. IQR \$600.  $Q_1$  \$400. SD \$400.  
 d) Minimum \$330. Mean \$770. Median \$550. Range \$1320. IQR \$660.  $Q_1$  \$385. SD \$440.  
 5. Lowest score = 910. Mean = 1230. SD = 120.  $Q_3$  = 1350. Median = 1270. IQR = 240.  
 7. Your score was 2.2 standard deviations higher than the mean score in the class.  
 9. 65  
 11. In January, a high of 55 is not quite 2 standard deviations above the mean, whereas in July a high of 55 is more than 2 standard deviations lower than the mean. So it's less likely to happen in July.

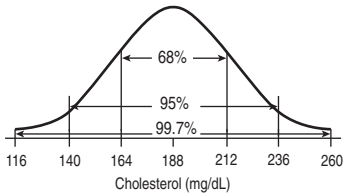
- 13. The z-scores, which account for the difference in the distributions of the two tests, are 1.5 and 0 for Derrick and 0.5 and 2 for Julie. Derrick's total is 1.5, which is less than Julie's 2.5.
- 15. a) Megan b) Anna
- 17. a) About 1.81 standard deviations below the mean.  
b) 1000 ( $z = 1.81$ ) is more unusual than 1250 ( $z = 1.17$ ).
- 19. a) Mean =  $1152 - 1000 = 152$  pounds; SD is unchanged at 84 pounds.  
b) Mean =  $0.40(1152) = \$460.80$ ; SD =  $0.40(84) = \$33.60$ .
- 21. Min =  $0.40(980) - 20 = \$372$ ;  
median =  $0.40(1140) - 20 = \$436$ ;  
SD =  $0.40(84) = \$33.60$ ; IQR =  $0.40(102) = \$40.80$ .
- 23. College professors can have between 0 and maybe 40 (or possibly 50) years' experience. A standard deviation of 1/2 year is impossible, because many professors would be 10 or 20 SDs away from the mean, whatever it is. An SD of 16 years would mean that 2 SDs on either side of the mean is plus or minus 32, for a range of 64 years. That's too high. So, the SD must be 6 years.
- 25. a)



- b) 18.6 to 31.0 mpg c) 16%
- d) 13.5% e) less than 12.4 mpg
- 27. Any weight more than 2 standard deviations below the mean, or less than  $1152 - 2(84) = 984$  pounds, is unusually low. We expect to see a steer below  $1152 - 3(84) = 900$  pounds only rarely.
- 29. a)



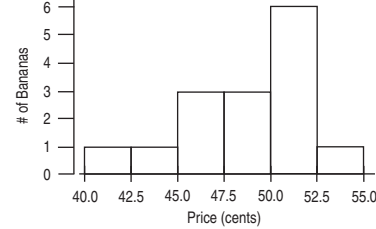
- b) Between 1.0 and 19.8 inches c) 2.5%
- d) 34% e) 16%
- 31. Since the histogram is not unimodal and symmetric, it is not wise to have faith in numbers from the Normal model.
- 33. a) 16% b) 3.8%  
c) Because the Normal model doesn't fit well.  
d) Distribution is skewed to the right.
- 35. a) 2.5%  
b) 2.5% of the receivers should gain less than -333 yards, but that's impossible, so the model doesn't fit well.  
c) Data are strongly skewed to the right, not symmetric.
- 37. a) 12.2% b) 71.6% c) 23.3%
- 39. a) 1259.7 lb b) 1081.3 lb c) 1108 lb to 1196 lb
- 41. a) 1130.7 lb b) 1347.4 lb c) 113.3 lb
- 43. a)



- b) 30.85% c) 17.00% d) 32 points e) 212.9 points
- 45. a) 11.1% b) (35.9, 40.5) inches c) 40.5 inches
- 47. a) 5.3 grams b) 6.4 grams  
c) Younger because SD is smaller.

PART I REVIEW

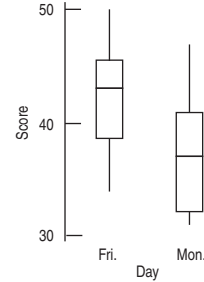
1. a)



- b) Median 49 cents, IQR 6 cents.
- c) The distribution is unimodal and left skewed. The center is near 50 cents; values range from 42 cents to 53 cents.
- 3. a) If enough sopranos have a height of 65 inches, this can happen.
- b) The distribution of heights for each voice part is roughly symmetric. The basses are slightly taller than the tenors. The sopranos and altos have about the same median height. Heights of basses and sopranos are more consistent than those of altos and tenors.

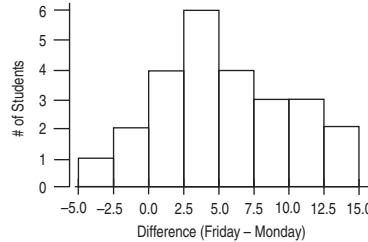
- 5. a) It means their heights are also more variable.
- b) The z-score for women to qualify is 2.40, compared with 1.75 for men, so it is harder for women to qualify.
- 7. a) *Who*—People who live near State University  
*What*—Age, attended college? Favorable opinion of State?  
*When*—Not stated  
*Where*—Region around State U.  
*Why*—To report to the university's directors  
*How*—Sampled and phoned 850 local residents
- b) Age—Quantitative (years); attended college?—categorical; favorable opinion?—categorical.
- c) The fact that the respondents know they are being interviewed by the university's staff may influence answers.
- 9. a) These are categorical data, so mean and standard deviation are meaningless.  
b) Not appropriate. Even if it fits well, the Normal model is meaningless for categorical data.

11. a)



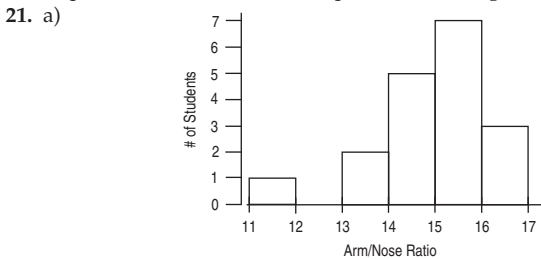
- b) The scores on Friday were higher by about 5 points on average. This is a drop of more than 10% off the average score and shows that students fared worse on Monday after preparing for the test on Friday. The spreads are about the same, but the scores on Monday are a bit skewed to the right.

c)



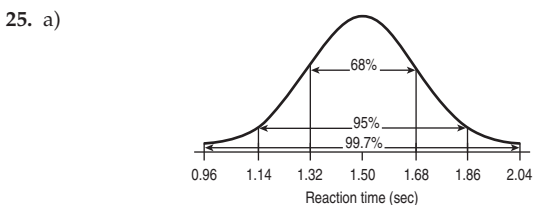
- d) The changes (Friday-Monday) are unimodal and centered near 4 points, with a spread of about 5 (SD). They are fairly symmetric, but slightly skewed to the right. Only 3 students did better on Monday (had a negative difference).

13. a) Categorical  
 b) Go fish. All you need to do is match the denomination. The denominations are not ordered. (Answers will vary.)  
 c) Gin rummy. All cards are worth their value in points (face cards are 10 points). (Answers will vary.)
15. a) Annual mortality rate for males (quantitative) in deaths per 100,000 and water hardness (quantitative) in parts per million.  
 b) Calcium is skewed right, possibly bimodal. There looks to be a mode down near 12 ppm that is the center of a fairly tight symmetric distribution and another mode near 62.5 ppm that is the center of a much more spread out, symmetric (almost uniform) distribution. Mortality, however, appears unimodal and symmetric with the mode near 1500 deaths per 100,000.
17. a) They are on different scales.  
 b) January's values are lower and more spread out.  
 c) Roughly symmetric but slightly skewed to the left. There are more low outliers than high ones. Center is around 40 degrees with an IQR of around 7.5 degrees.
19. a) Bimodal with modes near 2 and 4.5 minutes. Fairly symmetric around each mode.  
 b) Because there are two modes, which probably correspond to two different groups of eruptions, an average might not make sense.  
 c) The intervals between eruptions are longer for long eruptions. There is very little overlap. More than 75% of the short eruptions had intervals less than about an hour (62.5 minutes), while more than 75% of the long eruptions had intervals longer than about 75 minutes. Perhaps the interval could even be used to predict whether the next eruption will be long or short.



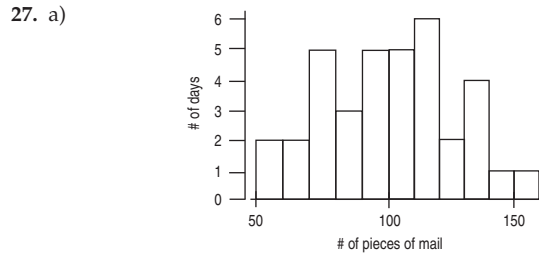
The distribution is left skewed with a center of about 15. It has an outlier between 11 and 12.

- b) Even though the distribution is somewhat skewed, the mean and median are close. The mean is 15.0 and the SD is 1.25.  
 c) Yes. 11.8 is already an outlier. 9.3 is more than 4.5 SDs below the mean. It is a very low outlier.
23. If we look only at the overall statistics, it appears that the follow-up group is insured at a much lower rate than those not traced (11.1% of the time compared with 16.6%). But most of the follow-up group were black, who have a lower rate of being insured. When broken down by race, the follow-up group actually has a higher rate of being insured for both blacks and whites. So the overall statistic is misleading and is attributable to the difference in race makeup of the two groups.

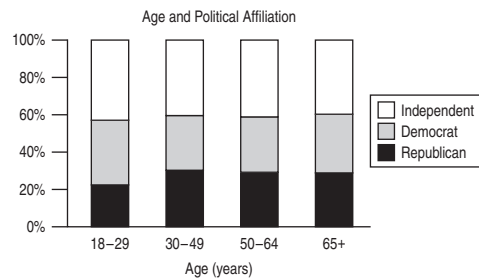


- b) According to the model, reaction times are symmetric with center at 1.5 seconds. About 95% of all reaction times are between 1.14 and 1.86 seconds.

- c) 8.2%  
 d) 24.1%  
 e) Quartiles are 1.38 and 1.62 seconds, so the IQR is 0.24 seconds.  
 f) The slowest 1/3 of all drivers have reaction times of 1.58 seconds or more.



- b) Mean 100.25, SD 25.54 pieces of mail.  
 c) The distribution is somewhat symmetric and unimodal, but the center is rather flat, almost uniform.  
 d) 64%. The Normal model seems to work reasonably well, since it predicts 68%.
29. a) *Who*—100 health food store customers  
*What*—Have you taken a cold remedy?, and Effectiveness (scale 1 to 10)  
*When*—Not stated  
*Where*—Not stated  
*Why*—Promotion of herbal medicine  
*How*—In-person interviews  
 b) Have you taken a cold remedy?—categorical. Effectiveness—categorical or ordinal.  
 c) No. Customers are not necessarily representative, and the Council had an interest in promoting the herbal remedy.
31. a) 38 cars  
 b) Possibly because the distribution is skewed to the right.  
 c) Center—median is 148.5 cubic inches. Spread—IQR is 126 cubic inches.  
 d) No. It's bigger than average, but smaller than more than 25% of cars. The upper quartile is at 231 inches.  
 e) No. 1.5 IQR is 189, and  $105 - 189$  is negative, so there can't be any low outliers.  $231 + 189 = 420$ . There aren't any cars with engines bigger than this, since the maximum has to be at most  $105$  (the lower quartile) +  $275$  (the range) =  $380$ .  
 f) Because the distribution is skewed to the right, this is probably not a good approximation.  
 g) Mean, median, range, quartiles, IQR, and SD all get multiplied by 16.4.
33. a) 30.4%  
 b) If this were a random sample of all voters, yes.  
 c) 36.6%  
 d) 8.8%  
 e) 23.1%  
 f) 47.0%
35. a) Republican—16,535, Democrat—17,183, Other—20,666; or Republican—30.4%, Democrat—31.6%, Other—38.0%.



- c) Among voters over 30, political affiliation appears to be largely unrelated to age. However there is some evidence that younger voters are less likely to be Republican  
 d) Voters who identified themselves as "Other" seem to be generally younger than Democrats or Republicans.



37. a) 0.43 hours. b) 1.4 hours.  
 c) 0.89 hours (or 53.4 minutes).  
 d) Survey results vary, and the mean and the SD may have changed.

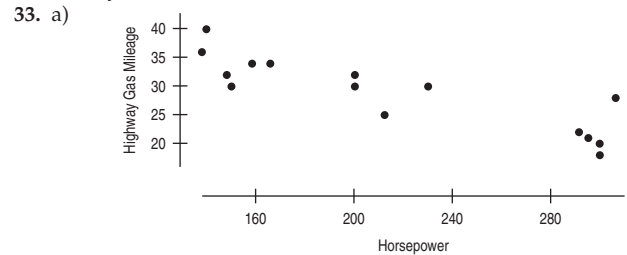
**CHAPTER 7**

1. a) Weight in ounces: explanatory; Weight in grams: response. (Could be other way around.) To predict the weight in grams based on ounces. Scatterplot: positive, straight, strong (perfectly linear relationship).  
 b) Circumference: explanatory. Weight: response. To predict the weight based on the circumference. Scatterplot: positive, linear, moderately strong.  
 c) Shoe size: explanatory; GPA: response. To try to predict GPA from shoe size. Scatterplot: no direction, no form, very weak.  
 d) Miles driven: explanatory; Gallons remaining: response. To predict the gallons remaining in the tank based on the miles driven since filling up. Scatterplot: negative, straight, moderate.
3. a) Altitude: explanatory; Temperature: response. (Other way around possible as well.) To predict the temperature based on the altitude. Scatterplot: negative, possibly straight, weak to moderate.  
 b) Ice cream cone sales: explanatory. Air-conditioner sales: response—although the other direction would work as well. To predict one from the other. Scatterplot: positive, straight, moderate.  
 c) Age: explanatory; Grip strength: response. To predict the grip strength based on age. Scatterplot: curved down, moderate. Very young and elderly would have grip strength less than that of adults.  
 d) Reaction time: explanatory; Blood alcohol level: response. To predict blood alcohol level from reaction time test. (Other way around is possible.) Scatterplot: positive, nonlinear, moderately strong.
5. a) None b) 3 and 4 c) 2, 3, and 4  
 d) 1 and 2 e) 3 and possibly 1
7. There seems to be a very weak—or possibly no—relation between brain size and performance IQ.
9. a)

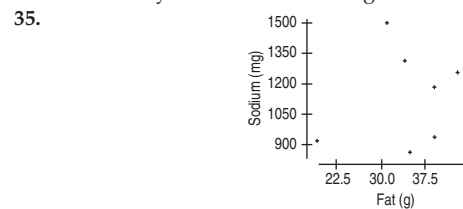


- b) Unimodal, skewed to the right. The skew.  
 c) The positive, somewhat linear relation between batch number and broken pieces.
11. a) 0.006 b) 0.777 c)  $-0.923$  d)  $-0.487$
13. There may be an association, but not a correlation unless the variables are quantitative. There could be a correlation between average number of hours of TV watched per week per person and number of crimes committed per year. Even if there is a relationship, it doesn't mean one causes the other.
15. a) Yes. It shows a linear form and no outliers.  
 b) There is a strong, positive, linear association between drop and speed; the greater the coaster's initial drop, the higher the top speed.
17. The scatterplot is not linear; correlation is not appropriate.
19. The correlation may be near 0. We expect nighttime temperatures to be low in January, increase through spring and into the summer months, then decrease again in the fall and winter. The relationship is not linear.

21. The correlation coefficient won't change, because it's based on z-scores. The z-scores of the prediction errors are the same whether they are expressed in nautical miles or miles.
23. a) Assuming the relation is linear, a correlation of  $-0.772$  shows a strong relation in a negative direction.  
 b) Continent is a categorical variable. Correlation does not apply.
25. a) Actually, yes, taller children will tend to have higher reading scores, but this doesn't imply causation.  
 b) Older children are generally both taller and are better readers. Age is the lurking variable.
27. a) No. We don't know this from the correlation alone. There may be a nonlinear relationship or outliers.  
 b) No. We can't tell from the correlation what the form of the relationship is.  
 c) No. We don't know from the correlation coefficient.  
 d) Yes, the correlation doesn't depend on the units used to measure the variables.
29. This is categorical data even though it is represented by numbers. The correlation is meaningless.
31. a) The association is positive, moderately strong, and roughly straight, with several states whose HCI seems high for their median income and one state whose HCI appears low given its median income.  
 b) The correlation would still be 0.65.  
 c) The correlation wouldn't change.  
 d) DC would be a moderate outlier whose HCI is high for its median income. It would lower the correlation slightly.  
 e) No. We can only say that higher median incomes are associated with higher housing costs, but we don't know why. There may be other economic variables at work.

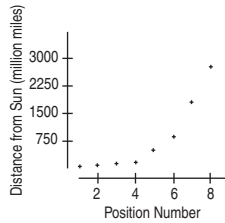


- b) Negative, linear, strong. c)  $-0.869$   
 d) There is a strong linear relation in a negative direction between horsepower and highway gas mileage. Lower fuel efficiency is associated with higher horsepower.



- (Plot could have explanatory and predictor variables swapped.) Correlation is 0.199. There does not appear to be a relation between sodium and fat content in burgers, especially without the low-fat, low-sodium item. The correlation of 0.199 shows a weak relationship, even with the outlier included.
37. a) Yes, the scatterplot appears to be somewhat linear.  
 b) As the number of runs increases, the attendance also increases.  
 c) There is a positive association, but it does not *prove* that more fans will come if the number of runs increases. Association does not indicate causality.
39. A scatterplot shows a generally straight scattered pattern with no outliers. The correlation between *Drop* and *Duration* is 0.35, indicating that rides on coasters with greater initial drops generally last somewhat longer, but the association is weak.

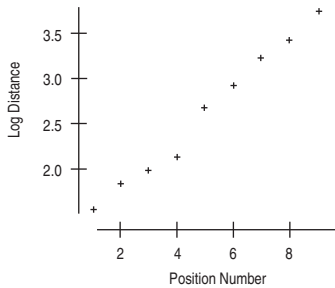
41. a)



The relation between position and distance is nonlinear, with a positive direction. There is very little scatter from the trend.

b) The relation is not linear.

c)



The relation between position number and log of distance appears to be roughly linear.

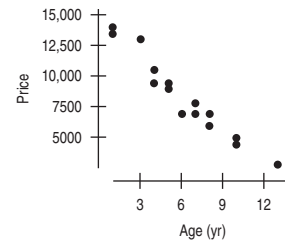
### CHAPTER 8

1. 281 milligrams
3. The potassium content is actually lower than the model predicts for a cereal with that much fiber.
5. The model predicts that cereals will have approximately 27 more milligrams of potassium for every additional gram of fiber.
7. 81.5%
9. The true potassium contents of cereals vary from the predicted amounts with a standard deviation of 30.77 milligrams.
11. a) Model is appropriate.  
b) Model is not appropriate. Relationship is nonlinear.  
c) Model may not be appropriate. Spread is changing.
13. 300 pounds/foot. It's ridiculous to suggest an extra foot in length would add 3, 30, or 3000 pounds to a car's weight.
15. a) *Price* (in thousands of dollars) is  $y$  and *Size* (in square feet) is  $x$ .  
b) Slope is thousands of \$ per square foot.  
c) Positive. Larger homes should cost more.
17. A linear model on *Size* accounts for 71.4% of the variation in home *Price*.
19. a) 0.845; + because larger homes cost more.  
b) Price should be 0.845 SDs above the mean in price.  
c) Price should be 1.690 SDs below the mean in price.
21. a) *Price* increases by about  $\$0.061 \times 1000$ , or \$61.00, per additional sq ft.  
b) 230.82 thousand, or \$230,820.  
c) \$115,020; \$6000 is the residual.
23. a)  $R^2$  does not tell whether the model is appropriate, but measures the strength of the linear relationship. High  $R^2$  could also be due to an outlier.  
b) Predictions based on a regression line are estimates of average values of  $y$  for a given  $x$ . The actual wingspan will vary around the prediction.
25. a) Probably not. Your score is better than about 97.5% of people, assuming scores follow the Normal model. Your next score is likely to be closer to the mean.  
b) The friend should probably retake the test. His score is better than only about 16% of people. His score is likely to be closer to the mean.
27. a) Probably. The residuals show some initially low points, but there is no clear curvature.

b) The linear model on *Tar* content accounts for 92.4% of the variability in *Nicotine*.

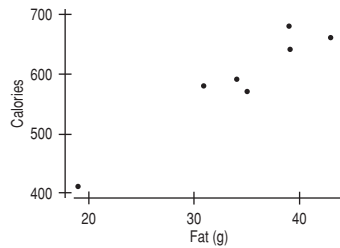
29. a)  $r = 0.961$   
b) Nicotine should be 1.922 SDs below average.  
c) *Tar* should be 0.961 SDs above average.
31. a)  $\widehat{Nicotine} = 0.15403 + 0.065052 Tar$   
b) 0.414 mg  
c) Predicted nicotine content increases by 0.065 mg of nicotine per additional milligram of tar.  
d) We'd expect a cigarette with no tar to have 0.154 mg of nicotine.  
e) 0.1094 mg
33. a) Yes. The relationship is straight enough, with a few outliers. The spread increases a bit for states with large median incomes, but we can still fit a regression line.  
b) From summary statistics:  $\widehat{HCI} = -156.50 + 0.0107 MFI$ ; from original data:  $\widehat{HCI} = -157.64 + 0.0107 MFI$   
c) From summary statistics: predicted HCI = 324.93; from original data: 324.87.  
d) 223.09 e)  $\widehat{z}_{HCI} = 0.65z_{MFI}$  f)  $\widehat{z}_{MFI} = 0.65z_{HCI}$
35. a)  $\widehat{Total} = 539.803 + 1.103Age$   
b) Yes. Both variables are quantitative; the plot is straight (although flat); there are no apparent outliers; the plot does not appear to change spread throughout the range of *Age*.  
c) \$559.65; \$594.94  
d) 0.14%  
e) No. The plot is nearly flat. The model explains almost none of the variation in *Total Yearly Purchases*.
37. a) Moderately strong, fairly straight, and positive. Possibly some outliers (higher-than-expected math scores).  
b) The student with 500 verbal and 800 math.  
c) Positive, fairly strong linear relationship. 46.9% of variation in math scores is explained by verbal scores.  
d)  $\widehat{Math} = 217.7 + 0.662 \times Verbal$ .  
e) Every point of verbal score adds 0.662 points to the predicted average math score.  
f) 548.5 points g) 53.0 points
39. a) 0.685 b)  $\widehat{Verbal} = 162.1 + 0.71 \times Math$ .  
c) The observed verbal score is higher than predicted from the math score  
d) 516.7 points. e) 559.6 points  
f) Regression to the mean. Someone whose math score is below average is predicted to have a verbal score below average, but not as far (in SDs). So if we use *that* verbal score to predict math, they will be even closer to the mean in predicted math score than their observed math score. If we kept cycling back and forth, eventually we would predict the mean of each and stay there.

41. a)

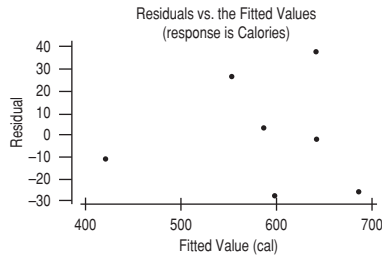


- b) Negative, linear, strong. c) Yes. d)  $-0.972$
- e) *Age* accounts for 94.4% of the variation in *Advertised Price*.
- f) Other factors contribute—options, condition, mileage, etc.
43. a)  $\widehat{Price} = 14,286 - 959 \times Years$ .  
b) Every extra year of age decreases average value by \$959.  
c) The average new Corolla costs a predicted \$14,286.  
d) \$7573  
e) Negative residual. Its price is below the predicted value for its age.  
f)  $-\$1195$   
g) No. After age 14, the model predicts negative prices. The relationship is no longer linear.

45. a)



- b) 92.3% of the variation in calories can be accounted for by the fat content.  
 c)  $\widehat{Calories} = 211.0 + 11.06 \times Fat$ .  
 d)



Residuals show no clear pattern, so the model seems appropriate.

- e) Could say a fat-free burger still has 211.0 calories, but this is extrapolation (no data close to 0).  
 f) Every gram of fat adds 11.06 calories, on average.  
 g) 553.5 calories.  
 47. a) The regression was for predicting calories from fat, not the other way around.  
 b)  $\widehat{Fat} = -15.0 + 0.083 \times Calories$ . Predict 34.8 grams of fat.  
 49. a)  $\% Body Fat = -27.4 + 0.25 \times Weight$ .  
 b) Residuals look randomly scattered around 0, so conditions are satisfied.  
 c)  $\% Body Fat$  increases, on average, by 0.25 percent per pound of  $Weight$ .  
 d) Reliable is relative.  $R^2$  is 48.5%, but residuals have a standard deviation of 7%, so variation around the line is large.  
 e) 0.9 percent.  
 51. a)  $\widehat{HighJump} = 2.681 - 0.00671 \times 800mTime$ . High-jump height is lower, on average, by 0.00671 meters per additional second of 800-m race time.  
 b) 16.4%  
 c) Yes, the slope is negative. Faster runners tend to jump higher.  
 d) There is a slight tendency for less variation in high-jump height among the slower runners than among the faster ones.  
 e) Not especially. The residual standard deviation is 0.060 meters, which is not much smaller than the SD of all high jumps (0.066 meters). The model doesn't appear to do a very good job of predicting.  
 53. The sum of the squared vertical distances to any other line would be greater than 1790.

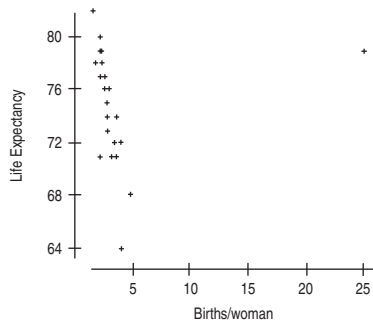
**CHAPTER 9**

1. a) The trend appears to be somewhat linear up to about 1940, but from 1940 to about 1970 the trend appears to be nonlinear. From 1975 or so to the present, the trend appears to be linear.  
 b) Relatively strong for certain periods.  
 c) No, as a whole the graph is clearly nonlinear. Within certain periods (ex: 1975 to the present) the correlation is high.  
 d) Overall, no. You could fit a linear model to the period from 1975 to 2003, but why? You don't need to interpolate, since every year is reported, and extrapolation seems dangerous.

3. a) The relationship is not straight.  
 b) It will be curved downward.  
 c) No. The relationship will still be curved.  
 5. a) No. We need to see the scatterplot first to see if the conditions are satisfied, and models are always wrong.  
 b) No, the linear model might not fit the data everywhere.  
 7. a) Millions of dollars per minute of run time.  
 b) Costs for movies increase at the same rate per minute.  
 c) On average dramas cost about \$20 million less for the same runtime.  
 9. a) The use of the Oakland airport has been growing at about 59,700 passengers/year, starting from about 282,000 in 1990.  
 b) 71% of the variation in passengers is accounted for by this model.  
 c) Errors in predictions based on this model have a standard deviation of 104,330 passengers.  
 d) No, that would extrapolate too far from the years we've observed.  
 e) The negative residual is September 2001. Air traffic was artificially low following the attacks on 9/11.  
 11. a) 1) High leverage, small residual.  
 2) No, not influential for the slope.  
 3) Correlation would decrease because outlier has large  $z_x$  and  $z_y$ , increasing correlation.  
 4) Slope wouldn't change much because the outlier is in line with other points.  
 b) 1) High leverage, probably small residual.  
 2) Yes, influential.  
 3) Correlation would weaken, increasing toward zero.  
 4) Slope would increase toward 0, since outlier makes it negative.  
 c) 1) Some leverage, large residual.  
 2) Yes, somewhat influential.  
 3) Correlation would increase, since scatter would decrease.  
 4) Slope would increase slightly.  
 d) 1) Little leverage, large residual.  
 2) No, not influential.  
 3) Correlation would become stronger and become more negative because scatter would decrease.  
 4) Slope would change very little.  
 13. 1) e 2) d 3) c 4) b 5) a  
 15. Perhaps high blood pressure causes high body fat, high body fat causes high blood pressure, or both could be caused by a lurking variable such as a genetic or lifestyle issue.  
 17. a) The graph shows that, on average, students progress at about one reading level per year. This graph shows averages for each grade. The linear trend has been enhanced by using averages.  
 b) Very close to 1.  
 c) The individual data points would show much more scatter, and the correlation would be lower.  
 d) A slope of 1 would indicate that for each 1-year grade level increase, the average reading level is increasing by 1 year.  
 19. a)  $Cost$  decreases by \$2.13 per degree of average daily  $Temp$ . So warmer temperatures indicate lower costs.  
 b) For an avg. monthly temperature of 0°F, the cost is predicted to be \$133.  
 c) Too high; the residuals (observed - predicted) around 32°F are negative, showing that the model overestimates the costs.  
 d) \$111.70 e) About \$105.70  
 f) No, the residuals show a definite curved pattern. The data are probably not linear.  
 g) No, there would be no difference. The relationship does not depend on the units.  
 21. a) 0.88  
 b) Interest rates during this period grew at about 0.25% per year, starting from an interest rate of about 0.64%.  
 c) Substituting 50 in the model yields a prediction of about 13%.  
 d) Not really. Extrapolating 20 years beyond the end of these data would be dangerous and unlikely to be accurate.

23. a) The two models fit comparably well, but they have very different slopes.  
 b) This model predicts the interest rate in 2000 to be 3.24%, much lower than the other model predicts.  
 c) We can trust the new predicted value because it is in the middle of the data used for the regression.  
 d) The best answer is "I can't predict that."
25. a) Stronger. Both slope and correlation would increase.  
 b) Restricting the study to nonhuman animals would justify it.  
 c) Moderately strong.  
 d) For every year increase in life expectancy, the gestation period increases by about 15.5 days, on average.  
 e) About 270.5 days.
27. a) Removing hippos would make the association stronger, since hippos are more of a departure from the pattern.  
 b) Increase.  
 c) No, there must be a good reason for removing data points.  
 d) Yes, removing it lowered the slope from 15.5 to 11.6 days per year.
29. a) Answers may vary. Using the data for 1955–2000 results in a scatterplot that is relatively linear with some curvature. The residuals plot shows a definite trend, indicating that the data are not linear. If you used the line, for 2010 the predicted age is 26.07 years.  
 b) Not much, since the data are not truly linear and 2010 is 10 years from the last data point (extrapolating is risky).  
 c) No, that extrapolation of more than 50 years would be absurd. There's no reason to believe the trend from 1955 to 2000 will continue.

31.

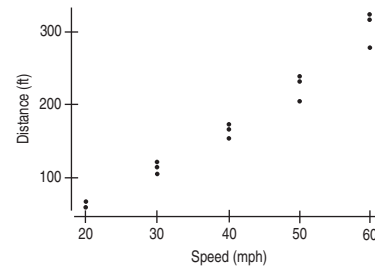


- a) Except for the outlier, Costa Rica, the data appear to have a linear form in a negative direction.  
 b) The outlier is Costa Rica, whose data appear to be wrong, with 25 births per woman. That's impossible.  
 c) With Costa Rica,  $r = 0.168$  and  $R\text{-squared} = 2.8\%$ , indicating that 2.8% of the variation in *Life Expectancy* is explained by the variation in *Births per Woman*. Without Costa Rica,  $r = -0.796$  and  $R\text{-squared} = 63.3\%$ , indicating that 63.3% of the variation in *Life Expectancy* is explained by the variation in *Births/Woman*.  
 d) With Costa Rica,  $\widehat{\text{Life Expectancy}} = 72.6 + 0.15 \text{ Births}$ ; without Costa Rica,  $\widehat{\text{Life Expectancy}} = 84.5 - 4.44 \text{ Births}$ .  
 e) The model with Costa Rica is not appropriate. The residuals plot shows a distinct outlier, which is Costa Rica. Removing Costa Rica gives a better residuals plot, suggesting that the linear equation is more appropriate.  
 f) With Costa Rica, the slope is near 0, suggesting that the linear model is not very useful. The  $y$ -intercept suggests that with no births, the life expectancy is about 72.6 years. Without Costa Rica, the slope is  $-4.44$ , indicating that an average increase of one child per woman predicts a lower life expectancy of 4.44 years, on average. The  $y$ -intercept indicates that a country with a birth rate of zero would have a life expectancy of 84.5 years. This is extrapolation.  
 g) While there is an association, there is no reason to expect causality. Lurking variables may be involved.

33. a) The scatterplot is clearly nonlinear; however, the last few years—say, from 1970 on—do appear to be linear.  
 b) Using the data from 1970 to 2006 gives  $r = 0.997$  and  $\widehat{\text{CPI}} = -9052.42 + 4.61 \text{ Year}$ . Predicted CPI in 2016 = 241.34 (an extrapolation of doubtful accuracy).

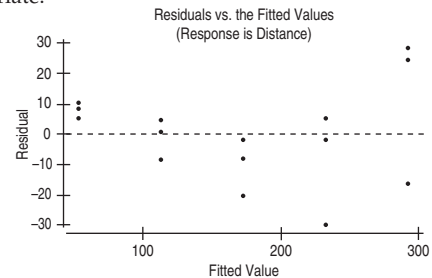
CHAPTER 10

1. a) No re-expression needed.  
 b) Re-express to straighten the relationship.  
 c) Re-express to equalize spread.
3. a) There's an annual pattern in when people fly, so the residuals cycle up and down.  
 b) No, this kind of pattern can't be helped by re-expression.
5. a) 16.44 b) 7.84 c) 0.36 d) 1.75 e) 27.59
7. a) Fairly linear, negative, strong.  
 b) Gas mileage decreases an average 7.652 mpg for each thousand pounds of weight.  
 c) No. Residuals show a curved pattern.
9. a) Residuals are more randomly spread around 0, with some low outliers.  
 b)  $\widehat{\text{Fuel Consumption}} = 0.625 + 1.178 \times \text{Weight}$ .  
 c) For each additional 1000 pounds of *Weight*, an additional 1.178 gallons will be needed to drive 100 miles.  
 d) 21.06 miles per gallon.
11. a) Although more than 97% of the variation in GDP can be accounted for by this model, we should examine a scatterplot of the residuals to see if it's appropriate.  
 b) No. The residuals show clear curvature.
13. Yes, the pattern in the residuals is somewhat weaker.
15. a)

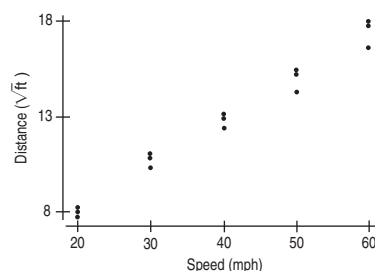


$$\widehat{\text{Distance}} = -65.9 + 5.98 \text{ Speed}$$

But residuals have a curved shape, so linear model is not appropriate.



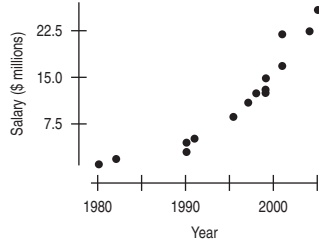
b)



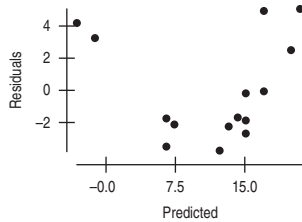
$\sqrt{\text{Distance}}$  linearizes the plot.

- c)  $\widehat{\text{Predicted } \sqrt{\text{Distance}}} = 3.30 + 0.235 \times \text{Speed}$ .  
 d) 263.4 feet. e) 390.2 feet (an extrapolation)

- f) Fairly confident, since  $R^2 = 98.4\%$ , and  $s$  is small.  
 17. a) The plot looks fairly straight. (It is okay to see a bend in the plot; there's one there.)



b)  $\widehat{Salary} = -1913.88 + 0.965 \text{ Year}$



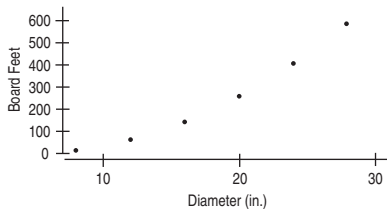
The residuals plot shows a strong bend.

- c)  $\log(\widehat{Salary})$  works well.  
 d)  $\log(\widehat{Salary}) = -109.133 + 0.05516 \text{ Year}$   
 19. a)
- 

$\log(\widehat{Distance})$  against position works pretty well.

$\log(\widehat{Distance}) = 1.245 + 0.271 \times \text{Position number}.$

- b) Pluto's residual is not especially larger in the log scale. However, a model without Pluto predicts the 9th planet should be 5741 million miles. Pluto, at "only" 3707 million miles, doesn't fit very well, giving support to the argument that Pluto doesn't behave like a planet.  
 21. The predicted  $\log(\widehat{Distance})$  of Eris is 3.685, corresponding to a distance of 4841 million miles. That's short of the actual average distance of 6300 million miles.  
 23. a)



$\widehat{\sqrt{Bdft}} = -4 + \text{diam}$   
 The model is exact.

- b) 36 board feet. c) 1024 board feet.  
 25.
- 

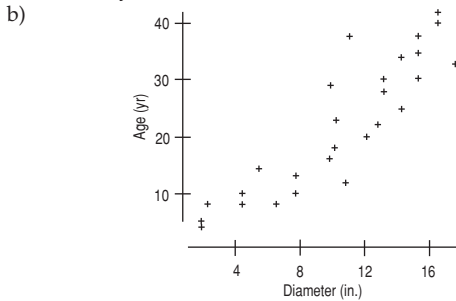
$\widehat{\log Life} = 1.685 + 0.18497 \log Decade$

27. The relationship cannot be made straight by the methods of this chapter.  
 29. a)  $\widehat{\sqrt{Left}} = 8.465 - 0.06926(\text{Age})$  b) 52.10 years  
 c) No; the residuals plot still shows a pattern.

**PART II REVIEW**

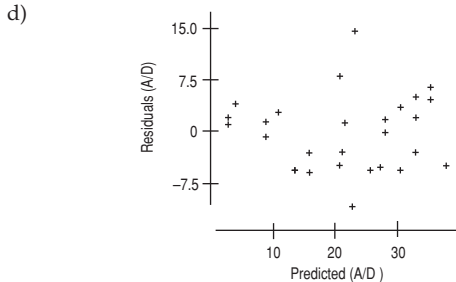
- % over 50, 0.69.  
 % under 20, -0.71.  
 % Graduating on time, -0.51.  
 % Full-time Faculty, 0.09
- a) There does not appear to be a linear relationship.  
 b) Nothing, there is no reason to believe that the results for the Finger Lakes region are representative of the vineyards of the world.  
 c)  $\widehat{CasePrice} = 92.77 + 0.567 \times \text{Years}.$   
 d) Only 2.7% of the variation in case price is accounted for by the ages of vineyards. Most of that is due to two outliers. We are better off using the mean price rather than this model.
- a)  $\widehat{TwinBirths} = -5119590 + 2618.25 \times \text{Year}.$   
 b) Each year, the number of twins born in a year increases, on average, by approximately 2618.25.  
 c) 143,092.5 births. The scatterplot appears to be somewhat linear, but there is some curvature in the pattern. There is no reason to believe that the increase will continue to be linear 5 years beyond the data.  
 d) The residuals plot shows a definite curved pattern, so the relation is not linear.
- a) -0.520  
 b) Negative, not strong, somewhat linear, but with more variation as pH increases.  
 c) The BCI would also be average.  
 d) The predicted BCI will be 1.56 SDs of BCI below the mean BCI.
- a)  $\widehat{ManateeDeaths} = -45.67 \times 0.1315 \text{ Powerboat Registrations (in 1000s)}.$   
 b) According to the model, for each increase of 10,000 motorboat registrations, the number of manatees killed increases by approximately 1.315.  
 c) If there were 0 motorboat registrations, the number of manatee deaths would be -45.67. This is obviously a silly extrapolation.  
 d) The predicted number is 82.41 deaths. The actual number of deaths was 79. The residual is  $79 - 82.41 = -3.41$ . The model overestimated the number of deaths by 3.41.  
 e) Negative residuals would suggest that the actual number of deaths was lower than the predicted number.  
 f) Over time, the number of motorboat registrations has increased and the number of manatee kills has increased. The trend may continue. Extrapolation is risky, however, because the government may enact legislation to protect the manatee.
- a) -0.984 b) 96.9% c) 32.95 mph d) 1.66 mph  
 e) Slope will increase.  
 f) Correlation will weaken (become less negative).  
 g) Correlation is the same, regardless of units.
- a) Weight (but unable to verify linearity).  
 b) As weight increases, mileage decreases.  
 c)  $\widehat{Weight}$  accounts for 81.5% of the variation in  $\widehat{Fuel Efficiency}.$
- a)  $\widehat{Horsepower} = 3.50 + 34.314 \times \text{Weight}.$   
 b) Thousands. For the equation to have predicted values between 60 and 160, the X values would have to be in thousands of pounds.  
 c) Yes. The residual plot does not show any pattern.  
 d) 115.0 horsepower.
- a) The scatterplot shows a fairly strong linear relation in a positive direction. There seem to be two distinct clusters of data.  
 b)  $\widehat{Interval} = 33.967 \div 10.358 \times \text{Duration}.$   
 c) The time between eruptions increases by about 10.4 minutes per minute of  $\widehat{Duration}$  on average.

- d) Since 77% of the variation in *Interval* is accounted for by *Duration* and the error standard deviation is 6.16 minutes, the prediction will be relatively accurate.
- e) 75.4 minutes.
- f) A residual is the observed value minus the predicted value. So the residual =  $79 - 75.4 = 3.6$  minutes, indicating that the model underestimated the interval in this case.
19. a)  $r = 0.888$ . Although  $r$  is high, you must look at the scatterplot and verify that the relation is linear in form.



The association between diameter and age appears to be strong, somewhat linear, and positive.

c)  $\widehat{Age} = -0.97 + 2.21 \times Diameter$ .



The residuals show a curved pattern (and two outliers).

- e) The residuals for five of the seven largest trees (15 in. or larger) are positive, indicating that the predicted values underestimate the age.
21. Most houses have areas between 1000 and 5000 square feet. Increasing 1000 square feet would result in either  $1000(.008) = 8$  thousand dollars,  $1000(.08) = 80$  thousand dollars,  $1000(.8) = 800$  thousand dollars, or  $1000(8) = 8000$  thousand dollars. Only \$80,000 is reasonable, so the slope must be 0.08.
23. a) The model predicts % smoking from year, not the other way around.
- b)  $\widehat{Year} = 2027.91 - 202.74 \times \% \text{ Smoking}$ .
- c) The smallest % smoking given is 12.7, and an extrapolation to  $x = 0$  is probably too far from the given data. The prediction is not very reliable in spite of the strong correlation.
25. The relation shows a negative direction, with a somewhat linear form, but perhaps with some slight curvature. There are several model outliers.
27. a) 71.9%
- b) As latitude increases, the January temperature decreases.
- c)  $\widehat{January \text{ Temperature}} = 108.80 - 2.111 \times Latitude$ .
- d) As the latitude increases by 1 degree, the average January temperature drops by about 2.11 degrees, on average.
- e) The  $y$ -intercept would indicate that the average January temperature is 108.8 when the latitude is 0. However, this is extrapolation and may not be meaningful.
- f) 24.4 degrees.
- g) The equation underestimates the average January temperature.
29. a) The scatterplot shows a strong, linear, positive association.
- b) There is an association, but it is likely that training and technique have increased over time and affected both jump performances.

- c) Neither; the change in units does not affect the correlation.
- d) The long-jumper would jump 0.925 SDs above the mean long jump, on average.
31. a) No relation; the correlation would probably be close to 0.
- b) The relation would have a positive direction and the correlation would be strong, assuming that students were studying French in each grade level. Otherwise, no correlation.
- c) No relation; correlation close to 0.
- d) The relation would have a positive direction and the correlation would be strong, since vocabulary would increase with each grade level.
33.  $\widehat{Calories} = 560.7 - 3.08 \times Time$ .  
Each minute extra at the table results in 3.08 fewer calories being consumed, on average. Perhaps the hungry children eat fast and eat more.
35. There seems to be a strong, positive, linear relationship with one high-leverage point (Northern Ireland) that makes the overall  $R^2$  quite low. Without that point, the  $R^2$  increases to 61.5%. Of course, these data are averaged across thousands of households, so the correlation appears to be higher than it would be for individuals. Any conclusions about individuals would be suspect.
37. a) 3.842    b) 501.187    c) 4.0
39. a) 30,818 pounds.  
b) 1302 pounds.  
c) 31,187.6 pounds.  
d) I would be concerned about using this relation if we needed accuracy closer than 1000 pounds or so, as the residuals are more than  $\pm 1000$  pounds.  
e) Negative residuals will be more of a problem, as the predicted weight would overestimate the weight of the truck; trucking companies might be inclined to take the ticket to court.
41. The original data are nonlinear, with a significant curvature. Using reciprocal square root of diameter gave a scatterplot that is nearly linear:  
$$1/\sqrt{\text{Drain Time}} = 0.0024 + 0.219 \text{ Diameter}$$

## CHAPTER 11

- Yes. You cannot predict the outcome beforehand.
- A machine pops up numbered balls. If it were truly random, the outcome could not be predicted and the outcomes would be equally likely. It is random only if the balls generate numbers in equal frequencies.
- Use two-digit numbers 00–99; let 00–02 = defect, 03–99 = no defect
- a) 45, 10    b) 17, 22
- If the lottery is random, it doesn't matter which number you play; all are equally likely to win.
- a) The outcomes are not equally likely; for example, tossing 5 heads does not have the same probability as tossing 0 or 9 heads, but the simulation assumes they are equally likely.  
b) The even-odd assignment assumes that the player is equally likely to score or miss the shot. In reality, the likelihood of making the shot depends on the player's skill.  
c) The likelihood for the first ace in the hand is not the same as for the second or third or fourth. But with this simulation, the likelihood is the same for each. (And it allows you to get 5 aces, which could get you in trouble in a real poker game!)
- The conclusion should indicate that the simulation suggests that the average length of the line would be 3.2 people. Future results might not match the simulated results exactly.
- a) The component is one voter voting. An outcome is a vote for our candidate or not. Use two random digits, giving 00–54 a vote for your candidate and 55–99 for the underdog.  
b) A trial is 100 votes. Examine 100 two-digit random numbers, and count how many people voted for each candidate. Whoever gets the majority of votes wins that trial.  
c) The response variable is whether the underdog wins or not.

17. Answers will vary, but average answer will be about 51%.
19. Answers will vary, but average answer will be about 26%.
21. a) Answers will vary, but you should win about 10% of the time.  
b) You should win at the same rate with any number.
23. Answers will vary, but you should win about 10% of the time.
25. Answers will vary, but average answer will be about 1.9 tests.
27. Answers will vary, but average answer will be about 1.24 points.
29. Do the simulation in two steps. First simulate the payoffs. Then count until \$500 is reached. Answers will vary, but average should be near 10.2 customers.
31. Answers will vary, but average answer will be about 3 children.
33. Answers will vary, but average answer will be about 7.5 rolls.
35. No, it will happen about 40% of the time.
37. Answers will vary, but average answer will be about 37.5%.
39. Three women will be selected about 7.8% of the time.

## CHAPTER 12

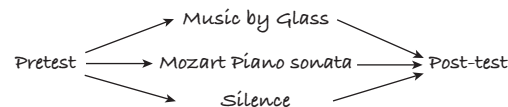
1. a) No. It would be nearly impossible to get exactly 500 males and 500 females from every country by random chance.  
b) A stratified sample, stratified by whether the respondent is male or female.
3. a) Voluntary response.  
b) We have no confidence at all in estimates from such studies.
5. a) The population of interest is all adults in the United States aged 18 and older.  
b) The sampling frame is U.S. adults with telephones.  
c) Some members of the population (e.g., many college students) don't have landline phones, which could create a bias.
7. a) Population—All U.S. adults.  
b) Parameter—Proportion who have used and benefited from alternative medicine.  
c) Sampling Frame—All Consumers Union subscribers.  
d) Sample—Those who responded.  
e) Method—Questionnaire to all (nonrandom).  
f) Bias—Nonresponse. Those who respond may have strong feelings one way or another.
9. a) Population—Adults.  
b) Parameter—Proportion who think drinking and driving is a serious problem.  
c) Sampling Frame—Bar patrons.  
d) Sample—Every 10th person leaving the bar.  
e) Method—Systematic sampling (may be random).  
f) Bias—Those interviewed had just left a bar. They may think drinking and driving is less of a problem than do other adults.
11. a) Population—Soil around a former waste dump.  
b) Parameter—Concentrations of toxic chemicals.  
c) Sampling Frame—Accessible soil around the dump.  
d) Sample—16 soil samples.  
e) Method—Not clear.  
f) Bias—Don't know if soil samples were randomly chosen. If not, may be biased toward more or less polluted soil.
13. a) Population—Snack food bags.  
b) Parameter—Weight of bags, proportion passing inspection.  
c) Sampling Frame—All bags produced each day.  
d) Sample—Bags in 10 randomly selected cases, 1 bag from each case for inspection.  
e) Method—Multistage random sampling.  
f) Bias—Should be unbiased.
15. Bias. Only people watching the news will respond, and their preference may differ from that of other voters. The sampling method may systematically produce samples that don't represent the population of interest.
17. a) Voluntary response. Only those who see the ad, have Internet access, and feel strongly enough will respond.  
b) Cluster sampling. One school may not be typical of all.

- c) Attempted census. Will have nonresponse bias.
- d) Stratified sampling with follow-up. Should be unbiased.
19. a) This is a multistage design, with a cluster sample at the first stage and a simple random sample for each cluster.  
b) If any of the three churches you pick at random is not representative of all churches, then you'll introduce sampling error by the choice of that church.
21. a) This is a systematic sample.  
b) The sampling frame is patrons willing to wait for the roller coaster on that day at that time. It should be representative of the people in line, but not of all people at the amusement park.  
c) It is likely to be representative of those waiting for the roller coaster. Indeed, it may do quite well if those at the front of the line respond differently (after their long wait) than those at the back of the line.
23. a) Answers will definitely differ. Question 1 will probably get many "No" answers, while Question 2 will get many "Yes" answers. This is response bias.  
b) "Do you think standardized tests are appropriate for deciding whether a student should be promoted to the next grade?" (Other answers will vary.)
25. a) Biased toward yes because of "pollute." "Should companies be responsible for any costs of environmental cleanup?"  
b) Biased toward no because of "old enough to serve in the military." "Do you think the drinking age should be lowered from 21?"
27. a) Not everyone has an equal chance. Misses people with unlisted numbers, or without landline phones, or at work.  
b) Generate random numbers and call at random times.  
c) Under the original plan, those families in which one person stays home are more likely to be included. Under the second plan, many more are included. People without landline phones are still excluded.  
d) It improves the chance of selected households being included.  
e) This takes care of phone numbers. Time of day may be an issue. People without landline phones are still excluded.
29. a) Answers will vary.  
b) Your own arm length. Parameter is your own arm length; population is all possible measurements of it.  
c) Population is now the arm lengths of you and your friends. The average estimates the mean of these lengths.  
d) Probably not. Friends are likely to be of the same age and not very diverse or representative of the larger population.
31. a) Assign numbers 001 to 120 to each order. Use random numbers to select 10 transactions to examine.  
b) Sample proportionately within each type. (Do a stratified random sample.)
33. a) Select three cases at random; then select one jar randomly from each case.  
b) Use random numbers to choose 3 cases from numbers 61 through 80; then use random numbers between 1 and 12 to select the jar from each case.  
c) No. Multistage sampling.
35. a) Depends on the Yellow Page listings used. If from regular (line) listings, this is fair if all doctors are listed. If from ads, probably not, as those doctors may not be typical.  
b) Not appropriate. This cluster sample will probably contain listings for only one or two business types.

## CHAPTER 13

1. a) No. There are no manipulated factors. Observational study.  
b) There may be lurking variables that are associated with both parental income and performance on the SAT.
3. a) This is a retrospective observational study.  
b) That's appropriate because MS is a relatively rare disease.

- c) The subjects were U.S. military personnel, some of whom had developed MS.  
d) The variables were the vitamin D blood levels and whether or not the subject developed MS.
5. a) This was a randomized, placebo-controlled experiment.  
b) Yes, such an experiment is the right way to determine whether black cohosh has an effect.  
c) 351 women aged 45 to 55 who reported at least two hot flashes a day.  
d) The treatments were black cohosh, a multiherb supplement, a multiherb supplement plus advice, estrogen, and a placebo. The response was the women's symptoms (presumably frequency of hot flashes).
7. a) Experiment.  
b) Bipolar disorder patients.  
c) Omega-3 fats from fish oil, two levels.  
d) 2 treatments.  
e) Improvement (fewer symptoms?).  
f) Design not specified.  
g) Blind (due to placebo), unknown if double-blind.  
h) Individuals with bipolar disease improve with high-dose omega-3 fats from fish oil.
9. a) Observational study.  
b) Prospective.  
c) Men and women with moderately high blood pressure and normal blood pressure, unknown selection process.  
d) Memory and reaction time.  
e) As there is no random assignment, there is no way to know that high blood pressure *caused* subjects to do worse on memory and reaction-time tests. A lurking variable may also be the cause.
11. a) Experiment.  
b) Postmenopausal women.  
c) Alcohol—2 levels; blocking variable—estrogen supplements (2 levels).  
d) 1 factor (alcohol) at 2 levels = 2 treatments.  
e) Increase in estrogen levels.  
f) Blocked.  
g) Not blind.  
h) Indicates that alcohol consumption *for those taking estrogen supplements* may increase estrogen levels.
13. a) Observational study.  
b) Retrospective.  
c) Women in Finland, unknown selection process with data from church records.  
d) Women's lifespans.  
e) As there is no random assignment, there is no way to know that having sons or daughters shortens or lengthens the lifespan of mothers.
15. a) Observational study.  
b) Prospective.  
c) People with or without depression, unknown selection process.  
d) Frequency of crying in response to sad situations.  
e) There is no apparent difference in crying response (to sad movies) for depressed and nondepressed groups.
17. a) Experiment.  
b) People experiencing migraines.  
c) 2 factors (pain reliever and water temperature), 2 levels each.  
d) 4 treatments.  
e) Level of pain relief.  
f) Completely randomized over 2 factors.  
g) Blind, as subjects did not know if they received the pain medication or the placebo, but not blind, as the subjects will know if they are drinking regular or ice water.  
h) It may indicate whether pain reliever alone or in combination with ice water gives pain relief, but patients are not blinded to ice water, so placebo effect may also be the cause of any relief seen caused by ice water.
19. a) Experiment.  
b) Athletes with hamstring injuries.  
c) 1 factor: type of exercise program (2 levels).  
d) 2 treatments.  
e) Time to return to sports.  
f) Completely randomized.  
g) No blinding—subjects must know what kind of exercise they do.  
h) Can determine which of the two exercise programs is more effective.
21. They need to compare omega-3 results to something. Perhaps bipolarity is seasonal and would have improved during the experiment anyway.
23. a) Subjects' responses might be related to many other factors (diet, exercise, genetics, etc). Randomization should equalize the two groups with respect to unknown factors.  
b) More subjects would minimize the impact of individual variability in the responses, but the experiment would become more costly and time consuming.
25. People who engage in regular exercise might differ from others with respect to bipolar disorder, and that additional variability could obscure the effectiveness of this treatment.
27. Answers may vary. Use a random-number generator to randomly select 24 numbers from 01 to 24 without replication. Assign the first 8 numbers to the first group, the second 8 numbers to the second group, and the third 8 numbers to the third group.
29. a) First, they are using athletes who have a vested interest in the success of the shoe by virtue of their sponsorship. They should choose other athletes. Second, they should randomize the order of the runs, not run all the races with their shoes second. They should blind the athletes by disguising the shoes if possible, so they don't know which is which. The timers shouldn't know which athletes are running with which shoes, either. Finally, they should replicate several times, since times will vary under both shoe conditions.  
b) Because of the problems in (a), the results they obtain may favor their shoes. In addition, the results obtained for Olympic athletes may not be the same as for the general runner.
31. a) Allowing athletes to self-select treatments could confound the results. Other issues such as severity of injury, diet, age, etc., could also affect time to heal; randomization should equalize the treatment groups with respect to any such variables.  
b) A control group could have revealed whether either exercise program was better (or worse) than just letting the injury heal.  
c) Doctors who evaluated the athletes to approve their return to sports should not know which treatment the subject had.  
d) It's hard to tell. The difference of 15 days seems large, but the standard deviations indicate that there was a great deal of variability in the times.
33. a) The differences among the Mozart and quiet groups were more than would have been expected from sampling variation.  
b)



- c) The Mozart group seems to have the smallest median difference and thus the *least* improvement, but there does not appear to be a significant difference.  
d) No, if anything, there is less improvement, but the difference does not seem significant compared with the usual variation.
35. a) Observational, prospective study.  
b) The supposed relation between health and wine consumption might be explained by the confounding variables of income and education.  
c) None of these. While the variables have a relation, there is no causality indicated for the relation.



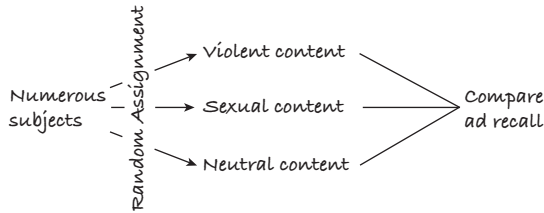
37. a) Arrange the 20 containers in 20 separate locations. Use a random-number generator to identify the 10 containers that should be filled with water.  
 b) Guessing, the dowser should be correct about 50% of the time. A record of 60% (12 out of 20) does not appear to be significantly different.  
 c) Answers may vary. You would need to see a high level of success—say, 90% to 100%, that is, 18 to 20 correct.
39. Randomly assign half the reading teachers in the district to use each method. Students should be randomly assigned to teachers as well. Make sure to block both by school and grade (or control grade by using only one grade). Construct an appropriate reading test to be used at the end of the year, and compare scores.
41. a) They mean that the difference is higher than they would expect from normal sampling variability.  
 b) An observational study.  
 c) No. Perhaps the differences are attributable to some confounding variable (e.g., people are more likely to engage in riskier behaviors on the weekend) rather than the day of admission.  
 d) Perhaps people have more serious accidents and traumas on weekends and are thus more likely to die as a result.
43. Answers may vary. This experiment has 1 factor (pesticide), at 3 levels (pesticide A, pesticide B, no pesticide), resulting in 3 treatments. The response variable is the number of beetle larvae found on each plant. Randomly select a third of the plots to be sprayed with pesticide A, a third with pesticide B, and a third with no pesticide (since the researcher also wants to know whether the pesticides even work at all). To control the experiment, the plots of land should be as similar as possible with regard to amount of sunlight, water, proximity to other plants, etc. If not, plots with similar characteristics should be blocked together. If possible, use some inert substance as a placebo pesticide on the control group, and do not tell the counters of the beetle larvae which plants have been treated with pesticides. After a given period of time, count the number of beetle larvae on each plant and compare the results.
- 
- ```

    graph LR
      A[Plots of corn] -- R --> B[Group 1 - pesticide A]
      A -- G --> C[Group 2 - pesticide B]
      A -- D --> D[Group 3 - no pesticide]
      B --> E[Count the number of beetle larvae on each plant and compare]
      C --> E
      D --> E
  
```
45. Answers may vary. Find a group of volunteers. Each volunteer will be required to shut off the machine with his or her left hand and right hand. Randomly assign the left or right hand to be used first. Complete the first attempt for the whole group. Now repeat the experiment with the alternate hand. Check the differences in time for the left and right hands.
47. a) Jumping with or without a parachute.  
 b) Volunteer skydivers (the dimwitted ones).  
 c) A parachute that looks real but doesn't work.  
 d) A good parachute and a placebo parachute.  
 e) Whether parachutist survives the jump (or extent of injuries).  
 f) All should jump from the same altitude in similar weather conditions and land on similar surfaces.  
 g) Randomly assign people the parachutes.  
 h) The skydivers (and the people involved in distributing the parachute packs) shouldn't know who got a working chute. And the people evaluating the subjects after the jumps should not be told who had a real parachute either!
3. Experiment, matched by gender and weight, randomization within blocks of two pups of same gender and weight. Factor: type of diet. Treatments: low-calorie diet and allowing the dog to eat all it wants. Response variable: length of life. Can conclude that, on average, dogs with a lower-calorie diet live longer.
5. Observational prospective study. Indicates folate *may* help in reducing colon cancer for those with family histories of the disease.
7. Sampling. Probably a simple random sample, although may be stratified by type of firework. Population is all fireworks produced each day. Parameter is proportion of duds. Can determine if the day's production is ready for sale.
9. Observational retrospective study. Living near strong electromagnetic fields may be associated with more leukemia than normal. May be lurking variables, such as socioeconomic level.
11. Experiment. Blocked by sex of rat. Randomization is not specified. Factor is type of hormone given. Treatments are leptin and insulin. Response variable is lost weight. Can conclude that hormones can help suppress appetites in rats, and the type of hormone varies by gender.
13. Experiment. Factor is gene therapy. Hamsters were randomized to treatments. Treatments were gene therapy or not. Response variable is heart muscle condition. Can conclude that gene therapy is beneficial (at least in hamsters).
15. Sampling. Population is all oranges on the truck. Parameter is proportion of unsuitable oranges. Procedure is probably simple random sampling. Can conclude whether or not to accept the truckload.
17. Observational prospective study. Physically fit men may have a lower risk of death from cancer.
19. Answers will vary. This is a simulation problem. Using a random digits table or software, call 0–4 a loss and 5–9 a win for the gambler on a game. Use blocks of 5 digits to simulate a week's pick.
21. Answers will vary.
23. a) Experiment. Actively manipulated candy giving, diners were randomly assigned treatments, control group was those with no candy, lots of dining parties.  
 b) It depends on when the decision was made. If early in the meal, the server may give better treatment to those who will receive candy—biasing the results.  
 c) A difference in response so large it cannot be attributed to natural sampling variability.
25. a) Voluntary response. Only those who feel strongly will pay for the 900 phone call.  
 b) "If it would help future generations live a longer, healthier life, would you be in favor of human cloning?"
27. a) Simulation results will vary. Average will be around 5.8 points.  
 b) Simulation results will vary. Average will also be around 5.8 points.  
 c) Answers will vary.
29. a) Yes.  
 b) No. Residences without phones are excluded. Residences with more than one phone had a higher chance.  
 c) No. People who respond to the survey may be of age but not registered voters.  
 d) No. Households who answered the phone may be more likely to have someone at home when the phone call was generated. These may not be representative of all households.
31. a) Does not prove it. There may be other confounding variables. Only way to prove this would be to do a controlled experiment.  
 b) Alzheimer's usually shows up late in life. Perhaps smokers have died of other causes before Alzheimer's can be seen.  
 c) An experiment would be unethical. One could design a prospective study in which groups of smokers and non-smokers are followed for many years and the incidence of Alzheimer's is tracked.

## PART III REVIEW

1. Observational prospective study. Indications of behavior differences can be seen in the two groups. May show a link between premature birth and behavior, but there may be lurking variables involved.

33.



Numerous subjects will be randomly assigned to see shows with violent, sexual, or neutral content. They will see the same commercials. After the show, they will be interviewed for their recall of brand names in the commercials.

35. a) May have been a simple random sample, but given the relative equality in age groups, may have been stratified.  
 b) 35.1%.  
 c) We don't know. Perhaps cell phones or unlisted numbers were excluded, and Democrats have more (or fewer) of those. Probably OK, though.  
 d) Do party affiliations differ for different age groups?
37. The factor in the experiment will be type of bird control. I will have three treatments: scarecrow, netting, and no control. I will randomly assign several different areas in the vineyard to one of the treatments, taking care that there is sufficient separation that the possible effect of the scarecrow will not be confounded. At the end of the season, the response variable will be the proportion of bird-damaged grapes.
39. a) We want all subjects treated as alike as possible. If there were no "placebo surgery," subjects would know this and perhaps behave differently.  
 b) The experiment looked for a difference in the effectiveness of the two treatments. (If we wanted to generalize, we would need to assume that the results for these volunteers are the same as on all patients who might need this operation.)  
 c) "Not statistically significant" means the difference in results were small enough that it could be explained by natural sampling variability.
41. a) Use stratified sampling to select 2 first-class passengers and 12 from coach.  
 b) Number passengers alphabetically, 01 = Bergman to 20 = Testut. Read in blocks of two, ignoring any numbers more than 20. This gives 65, 43, 67, 11 (selects Fontana), 27, 04 (selects Castillo).  
 c) Number passengers alphabetically from 001 to 120. Use the random-number table to find three-digit numbers in this range until 12 different values have been selected.
43. Simulation results will vary.  
 (Use integers 00 to 99 as a basis. Use integers 00 to 69 to represent a tee shot on the fairway. If on the fairway, use digits 00 to 79 to represent on the green. If off the fairway, use 00 to 39 to represent getting on the green. If not on the green, use digits 00 to 89 to represent landing on the green. For the first putt, use digits 00 to 19 to represent making the shot. For subsequent putts, use digits 00 to 89 to represent making the shot.)

**CHAPTER 14**

1. a)  $S = \{HH, HT, TH, TT\}$ , equally likely.  
 b)  $S = \{0, 1, 2, 3\}$ , not equally likely.  
 c)  $S = \{H, TH, TTH, TTT\}$ , not equally likely.  
 d)  $S = \{1, 2, 3, 4, 5, 6\}$ , not equally likely.
3. In this context "truly random" should mean that every number is equally likely to occur.
5. There is no "Law of Averages." She would be wrong to think that they are "due" for a harsh winter.
7. There is no "Law of Averages." If at bats are independent, his chance for a hit does not change based on recent successes or failures.

9. a) There is some chance you would have to pay out much more than the \$300.  
 b) Many customers pay for insurance. The small risk for any one customer is spread among all.
11. a) Legitimate. b) Legitimate.  
 c) Not legitimate (sum more than 1). d) Legitimate.  
 e) Not legitimate (can't have negatives or values more than 1).
13. A family may own both a car and an SUV. The events are not disjoint, so the Addition Rule does not apply.
15. When cars are traveling close together, their speeds are not independent, so the Multiplication Rule does not apply.
17. a) He has multiplied the two probabilities.  
 b) He assumes that being accepted at the colleges are independent events.  
 c) No. Colleges use similar criteria for acceptance, so the decisions are not independent.
19. a) 0.72 b) 0.89 c) 0.28
21. a) 0.5184 b) 0.0784 c) 0.4816
23. a) Repair needs for the two cars must be independent.  
 b) Maybe not. An owner may treat the two cars similarly, taking good (or poor) care of both. This may decrease (or increase) the likelihood that each needs to be repaired.
25. a)  $342/1005 = 0.340$ .  
 b)  $30/1005 + 50/1005 = 80/1005 = 0.080$ .
27. a) 0.195 b) 0.913  
 c) Responses are independent.  
 d) People were polled at random.
29. a) 0.4712 b) 0.7112  
 c)  $(1 - 0.76) + 0.76(1 - 0.38)$  or  $1 - (0.76)(0.38)$
31. a) 1) 0.30 2) 0.30 3) 0.90 4) 0.0  
 b) 1) 0.027 2) 0.128 3) 0.512 4) 0.271
33. a) Disjoint (can't be both red and orange).  
 b) Independent (unless you're drawing from a small bag).  
 c) No. Once you know that one of a pair of disjoint events has occurred, the other is impossible.
35. a) 0.0046 b) 0.125 c) 0.296 d) 0.421 e) 0.995
37. a) 0.027 b) 0.063 c) 0.973 d) 0.014
39. a) 0.024 b) 0.250 c) 0.543
41. 0.078.
43. a) For any day with a valid three-digit date, the chance is 0.001, or 1 in 1000. For many dates in October through December, the probability is 0. (No three digits will make 10/15, for example.)  
 b) There are 65 days when the chance to match is 0. (Oct. 10–31, Nov. 10–30, and Dec. 10–31.) The chance for no matches on the remaining 300 days is 0.741  
 c) 0.259 d) 0.049

**CHAPTER 15**

1. a) 0.68 b) 0.32 c) 0.04
3. a) 0.31 b) 0.48 c) 0.31
5. a) 0.2025 b) 0.6965 c) 0.2404 d) 0.0402
7. a) 0.50 b) 1.00 c) 0.077 d) 0.333
9. a) 0.11 b) 0.27 c) 0.407 d) 0.344
11. a) 0.011 b) 0.222 c) 0.054 d) 0.337 e) 0.436
13. 0.21
15. a) 0.145 b) 0.118 c) 0.414 d) 0.217
17. a) 0.318 b) 0.955 c) 0.071 d) 0.009
19. a) 32% b) 0.135  
 c) No, 7% of juniors have taken both.  
 d) No, the probability that a junior has taken a computer course is 0.23. The probability that a junior has taken a computer course given he or she has taken a Statistics course is 0.135.
21. a) 0.266  
 b) No, 26.6% of homes with garages have pools; 21% of homes overall have pools.  
 c) No, 17% of homes have both.

23. Yes,  $P(\text{Ace}) = 4/52$ .  $P(\text{Ace} | \text{any suit}) = 1/13$ .
25. a) 0.17  
 b) No; 13% of the chickens had both contaminants.  
 c) No;  $P(C|S) = 0.87 \neq P(C)$ . If a chicken is contaminated with salmonella, it's more likely also to have campylobacter.
27. No, only 32% of all men have high cholesterol, but 40.7% of those with high blood pressure do.
29. a) 95.6%  
 b) Probably. 95.4% of people with cell phones had landlines, and 95.6% of all people did.
31. No. Only 34% of men were Democrats, but over 41% of all voters were.
33. a) No, the probability that the luggage arrives on time depends on whether the flight is on time. The probability is 95% if the flight is on time and only 65% if not.  
 b) 0.695
35. 0.975
37. a) No, the probability of missing work for day-shift employees is 0.01. It is 0.02 for night-shift employees. The probability depends on whether they work day or night shift.  
 b) 1.4%
39. 57.1%
41. a) 0.20    b) 0.272    c) 0.353    d) 0.033
43. 0.563    45. Over 0.999

**CHAPTER 16**

1. a) 19                      b) 4.2
3. a) 

|                        |                 |                 |                 |                |
|------------------------|-----------------|-----------------|-----------------|----------------|
| Amount won             | \$0             | \$5             | \$10            | \$30           |
| $P(\text{Amount won})$ | $\frac{26}{52}$ | $\frac{13}{52}$ | $\frac{12}{52}$ | $\frac{1}{52}$ |
- b) \$4.13                      c) \$4 or less (answers may vary)
5. a) 

|                      |     |      |      |
|----------------------|-----|------|------|
| Children             | 1   | 2    | 3    |
| $P(\text{Children})$ | 0.5 | 0.25 | 0.25 |
- b) 1.75 children    c) 0.875 boys
- |                  |     |      |       |       |
|------------------|-----|------|-------|-------|
| Boys             | 0   | 1    | 2     | 3     |
| $P(\text{Boys})$ | 0.5 | 0.25 | 0.125 | 0.125 |
7. \$27,000
9. a) 7                      b) 1.89
11. \$5.44
13. 0.83
15. a) 1.7                      b) 0.9
17.  $\mu = 0.64, \sigma = 0.93$
19. a) \$50                      b) \$100
21. a) No. The probability of winning the second depends on the outcome of the first.  
 b) 0.42                      c) 0.08
- d) 

|                       |      |      |      |
|-----------------------|------|------|------|
| Games won             | 0    | 1    | 2    |
| $P(\text{Games won})$ | 0.42 | 0.50 | 0.08 |
- e)  $\mu = 0.66, \sigma = 0.62$
23. a) 

|                         |       |       |       |
|-------------------------|-------|-------|-------|
| Number good             | 0     | 1     | 2     |
| $P(\text{Number good})$ | 0.067 | 0.467 | 0.467 |
- b) 1.40                      c) 0.61
25. a)  $\mu = 30, \sigma = 6$     b)  $\mu = 26, \sigma = 5$     c)  $\mu = 30, \sigma = 5.39$   
 d)  $\mu = -10, \sigma = 5.39$     e)  $\mu = 20, \sigma = 2.83$

27. a)  $\mu = 240, \sigma = 12.80$     b)  $\mu = 140, \sigma = 24$   
 c)  $\mu = 720, \sigma = 34.18$     d)  $\mu = 60, \sigma = 39.40$   
 e)  $\mu = 600, \sigma = 22.63$
29. a) 1.8    b) 0.87  
 c) Cartons are independent of each other.
31.  $\mu = 13.6, \sigma = 2.55$  (assuming the hours are independent of each other).
33. a)  $\mu = 23.4, \sigma = 2.97$   
 b) We assume each truck gets tickets independently.
35. a) There will be many gains of \$150 with a few large losses.  
 b)  $\mu = \$300, \sigma = \$8485.28$   
 c)  $\mu = \$1,500,000, \sigma = \$600,000$   
 d) Yes. \$0 is 2.5 SDs below the mean for 10,000 policies.  
 e) Losses are independent of each other. A major catastrophe with many policies in an area would violate the assumption.
37. a) 1 oz                      b) 0.5 oz                      c) 0.023  
 d)  $\mu = 4 \text{ oz}, \sigma = 0.5 \text{ oz}$   
 e) 0.159  
 f)  $\mu = 12.3 \text{ oz}, \sigma = 0.54 \text{ oz}$
39. a) 12.2 oz    b) 0.51 oz                      c) 0.058
41. a)  $\mu = 200.57 \text{ sec}, \sigma = 0.46 \text{ sec}$   
 b) No,  $z = \frac{199.48 - 200.57}{0.461} = -2.36$ . There is only 0.009 probability of swimming that fast or faster.
43. a)  $A$  = price of a pound of apples;  $P$  = price of a pound of potatoes; Profit =  $100A + 50P - 2$   
 b) \$63.00                      c) \$20.62  
 d) Mean—no; SD—yes (independent sales prices).
45. a)  $\mu = 1920, \sigma = 48.99; P(T > 2000) = 0.051$   
 b)  $\mu = \$220, \sigma = 11.09$ ; No—\$300 is more than 7 SDs above the mean.  
 c)  $P(D - \frac{1}{2}C > 0) \approx 0.26$

**CHAPTER 17**

1. a) No. More than two outcomes are possible.  
 b) Yes, assuming the people are unrelated to each other.  
 c) No. The chance of a heart changes as cards are dealt so the trials are not independent.  
 d) No, 500 is more than 10% of 3000.  
 e) If packages in a case are independent of each other, yes.
3. a) Use single random digits. Let 0, 1 = Tiger. Count the number of random numbers until a 0 or 1 occurs.  
 c) Results will vary.  
 d)
- |        |     |      |       |       |       |       |       |       |          |
|--------|-----|------|-------|-------|-------|-------|-------|-------|----------|
| $x$    | 1   | 2    | 3     | 4     | 5     | 6     | 7     | 8     | $\geq 9$ |
| $P(x)$ | 0.2 | 0.16 | 0.128 | 0.102 | 0.082 | 0.066 | 0.052 | 0.042 | 0.168    |
5. a) Use single random digits. Let 0, 1 = Tiger. Examine random digits in groups of five, counting the number of 0's and 1's.  
 c) Results will vary.  
 d)
- |        |      |      |      |      |      |     |
|--------|------|------|------|------|------|-----|
| $x$    | 0    | 1    | 2    | 3    | 4    | 5   |
| $P(x)$ | 0.33 | 0.41 | 0.20 | 0.05 | 0.01 | 0.0 |
7. Departures from the same airport during a 2-hour interval may not be independent. All could be delayed by weather, for example.
9. a) 0.0819                      b) 0.0064                      c) 0.992
11. 5    13. 20 calls
15. a) 25    b) 0.185    c) 0.217    d) 0.693
17. a) 0.0745    b) 0.502    c) 0.211  
 d) 0.0166    e) 0.0179    f) 0.9987
19. a) 0.65    b) 0.75    c) 7.69 picks
21. a)  $\mu = 10.44, \sigma = 1.16$   
 b) i) 0.812    ii) 0.475    iii) 0.00193    iv) 0.998

23.  $\mu = 20.28, \sigma = 4.22$   
 25. a) 0.118      b) 0.324      c) 0.744      d) 0.580  
 27. a)  $\mu = 56, \sigma = 4.10$   
 b) Yes,  $np = 56 \geq 10, nq = 24 \geq 10$ , serves are independent.  
 c) In a match with 80 serves, approximately 68% of the time she will have between 51.9 and 60.1 good serves, approximately 95% of the time she will have between 47.8 and 64.2 good serves, and approximately 99.7% of the time she will have between 43.7 and 68.3 good serves.  
 d) Normal, approx.: 0.014; Binomial, exact: 0.016  
 29. a) Assuming apples fall and become blemished independently of each other, Binom(300, 0.06) is appropriate. Since  $np \geq 10$  and  $nq \geq 10$ ,  $N(18, 4.11)$  is also appropriate.  
 b) Normal, approx.: 0.072; Binomial, exact: 0.085  
 c) No, 50 is 7.8 SDs above the mean.  
 31. Normal, approx.: 0.053; Binomial, exact: 0.061  
 33. The mean number of sales should be 24 with SD 4.60. Ten sales is more than 3.0 SDs below the mean. He was probably misled.  
 35. a) 5      b) 0.066      c) 0.107      d)  $\mu = 24, \sigma = 2.19$   
 e) Normal, approx.: 0.819; Binomial, exact: 0.848  
 37.  $\mu = 20, \sigma = 4$ . I'd want *at least* 32 (3 SDs above the mean). (Answers will vary.)  
 39. Probably not. There's a more than 9% chance that he could hit 4 shots in a row, so he can expect this to happen nearly once in every 10 sets of 4 shots he takes. That does not seem unusual.  
 41. Yes. We'd expect him to make 22 shots, with a standard deviation of 3.15 shots. 32 shots is more than 3 standard deviations above the expected value, an unusually high rate of success.

PART IV REVIEW

1. a) 0.34      b) 0.27      c) 0.069  
 d) No, 2% of cars have both types of defects.  
 e) Of all cars with cosmetic defects, 6.9% have functional defects. Overall, 7.0% of cars have functional defects. The probabilities here are estimates, so these are probably close enough to say the defects are independent.  
 3. a)  $C = \text{Price to China}; F = \text{Price to France}; \text{Total} = 3C + 5F$   
 b)  $\mu = \$5500, \sigma = \$672.68$       c)  $\mu = \$500, \sigma = \$180.28$   
 d) Means—no. Standard deviations—yes; ticket prices must be independent of each other for different countries, but all tickets to the same country are at the same price.  
 5. a)  $\mu = -\$0.20, \sigma = \$1.89$       b)  $\mu = -\$0.40, \sigma = \$2.67$   
 7. a) 0.106      b) 0.651      c) 0.442  
 9. a) 0.590      b) 0.328      c) 0.00856  
 11. a)  $\mu = 15.2, \sigma = 3.70$       b) Yes,  $np \geq 10$  and  $nq \geq 10$   
 c) Normal, approx.: 0.080; Binomial, exact: 0.097  
 13. a) 0.0173      b) 0.591  
 c) Left: 960; right: 120; both: 120  
 d)  $\mu = 120, \sigma = 10.39$   
 e) About 68% chance of between 110 and 130; about 95% between 99 and 141; about 99.7% between 89 and 151.  
 15. a) Men's heights are more variable than women's.  
 b) Men (1.75 SD vs 2.4 SD for women)  
 c)  $M = \text{Man's height}; W = \text{Woman's height}; M - W$  is how much taller the man is.  
 d) 5.1"      e) 3.75"      f) 0.913  
 g) If independent, it should be about 91.3%. We are told 92%. This difference seems small and may be due to natural sampling variability.  
 17. a) The chance is  $1.6 \times 10^{-7}$ .      b) 0.952      c) 0.063  
 19. \$240  
 21. a) 0.717      b) 0.588  
 23. a)  $\mu = 100, \sigma = 8$       b)  $\mu = 1000, \sigma = 60$   
 c)  $\mu = 100, \sigma = 8.54$       d)  $\mu = -50, \sigma = 10$   
 e)  $\mu = 100, \sigma = 11.31$

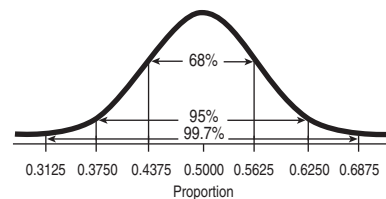
25. a) Many do both, so the two categories can total more than 100%.  
 b) No. They can't be disjoint. If they were, the total would be 100% or less.  
 c) No. Probabilities are different for boys and girls.  
 d) 0.0524  
 27. a) 21 days      b) 1649.73 som  
 c) 3300 som extra. About 157-som "cushion" each day.  
 29. No, you'd expect 541.2 homeowners, with an SD of 13.56. 523 is 1.34 SDs below the mean; not unusual.  
 31. a) 0.018      b) 0.300      c) 0.26  
 33. a) 6      b) 15      c) 0.402  
 35. a) 34%      b) 35%      c) 31.4%  
 d) 31.4% of classes that used calculators used computer assignments, while in classes that didn't use calculators, 30.6% used computer assignments. These are close enough to think the choice is probably independent.  
 37. a) 1/11      b) 7/22      c) 5/11      d) 0      e) 19/66  
 39. a) Expected number of stars with planets.  
 b) Expected number of planets with intelligent life.  
 c) Probability of a planet with a suitable environment having intelligent life.  
 d)  $f_i$ : If a planet has a suitable environment, the probability that life develops.  
 $f_j$ : If a planet develops life, the probability that the life evolves intelligence.  
 $f_c$ : If a planet has intelligent life, the probability that it develops radio communication.  
 41. 0.991

CHAPTER 18

1. All the histograms are centered near 0.05. As  $n$  gets larger, the histograms approach the Normal shape, and the variability in the sample proportions decreases.  
 3. a)

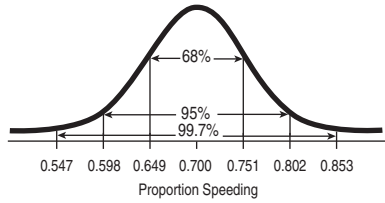
| $n$ | Observed mean | Theoretical mean | Observed st. dev. | Theoretical st. dev. |
|-----|---------------|------------------|-------------------|----------------------|
| 20  | 0.0497        | 0.05             | 0.0479            | 0.0487               |
| 50  | 0.0516        | 0.05             | 0.0309            | 0.0308               |
| 100 | 0.0497        | 0.05             | 0.0215            | 0.0218               |
| 200 | 0.0501        | 0.05             | 0.0152            | 0.0154               |

- b) They are all quite close to what we expect from the theory.  
 c) The histogram is unimodal and symmetric for  $n = 200$ .  
 d) The success/failure condition says that  $np$  and  $nq$  should both be at least 10, which is not satisfied until  $n = 200$  for  $p = 0.05$ . The theory predicted my choice.  
 5. a) Symmetric, because probability of heads and tails is equal.  
 b) 0.5      c) 0.125      d)  $np = 8 < 10$   
 7. a) About 68% should have proportions between 0.4 and 0.6, about 95% between 0.3 and 0.7, and about 99.7% between 0.2 and 0.8.  
 b)  $np = 12.5, nq = 12.5$ ; both are  $\geq 10$ .  
 c)



- $np = nq = 32$ ; both are  $\geq 10$ .  
 d) Becomes narrower (less spread around 0.5).  
 9. This is a fairly unusual result: about 2.26 SDs below the mean. The probability of that is about 0.012. So, in a class of 100 this is certainly a reasonable possibility.

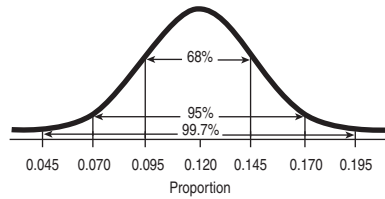
11. a)



b) Both  $np = 56$  and  $nq = 24 \geq 10$ . Drivers *may* be independent of each other, but if flow of traffic is very fast, they may not be. Or weather conditions may affect all drivers. In these cases they may get more or fewer speeders than they expect.

13. a) Assume that these children are typical of the population. They represent fewer than 10% of all children. We expect 20.4 near-sighted and 149.6 not; both are at least 10.

b)



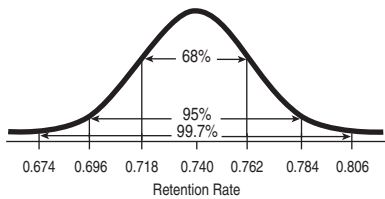
c) Probably between 12 and 29.

15. a)  $\mu = 7\%$ ,  $\sigma = 1.8\%$

b) Assume that clients pay independently of each other, that we have a random sample of all possible clients, and that these represent less than 10% of all possible clients.  $np = 14$  and  $nq = 186$  are both at least 10.

c) 0.048

17.



These are not random samples, and not all colleges may be typical (representative).  $np = 296$ ,  $nq = 104$  are both at least 10.

19. Yes; if their students were typical, a retention rate of  $522/603 = 86.6\%$  would be over 7 standard deviations above the expected rate of 74%.

21. 0.212. Reasonable that those polled are independent of each other and represent less than 10% of all potential voters. We assume the sample was selected at random. Success/Failure Condition met:  $np = 208$ ,  $nq = 192$ . Both  $\geq 10$ .

23. 0.088 using  $N(0.08, 0.022)$  model.

25. Answers will vary. Using  $\mu + 3\sigma$  for "very sure," the restaurant should have 89 nonsmoking seats. Assumes customers at any time are independent of each other, a random sample, and represent less than 10% of all potential customers.  $np = 72$ ,  $nq = 48$ , so Normal model is reasonable ( $\mu = 0.60$ ,  $\sigma = 0.045$ ).

27. a) Normal, center at  $\mu$ , standard deviation  $\sigma/\sqrt{n}$ .

b) Standard deviation will be smaller. Center will remain the same.

29. a) The histogram is unimodal and slightly skewed to the right, centered at 36 inches with a standard deviation near 4 inches.

b) All the histograms are centered near 36 inches. As  $n$  gets larger, the histograms approach the Normal shape and the variability in the sample means decreases. The histograms are fairly normal by the time the sample reaches size 5.

31. a)

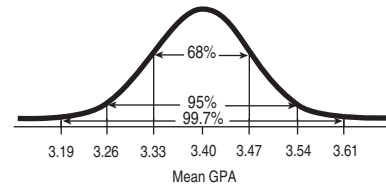
| $n$ | Observed mean | Theoretical mean | Observed st. dev. | Theoretical st. dev. |
|-----|---------------|------------------|-------------------|----------------------|
| 2   | 36.314        | 36.33            | 2.855             | 2.842                |
| 5   | 36.314        | 36.33            | 1.805             | 1.797                |
| 10  | 36.341        | 36.33            | 1.276             | 1.271                |
| 20  | 36.339        | 36.33            | 0.895             | 0.899                |

b) They are all very close to what we would expect.

c) For samples as small as 5, the sampling distribution of sample means is unimodal and very symmetric.

d) The distribution of the original data is nearly unimodal and symmetric, so it doesn't take a very large sample size for the distribution of sample means to be approximately Normal.

33.



Normal,  $\mu = 3.4$ ,  $\sigma = 0.07$ . We assume that the students are randomly assigned to the seminars and represent less than 10% of all possible students, and that individual's GPAs are independent of one another.

35. a) As the CLT predicts, there is more variability in the smaller outlets.

b) If the lottery is random, all outlets are equally likely to sell winning tickets.

37. a) 21.1%      b) 276.8 days or more

c)  $N(266, 2.07)$       d) 0.002

39. a) There are more premature births than very long pregnancies. Modern practice of medicine stops pregnancies at about 2 weeks past normal due date.

b) Parts (a) and (b)—yes—we can't use Normal model if it's very skewed. Part (c)—no—CLT guarantees a Normal model for this large sample size.

41. a)  $\mu = \$2.00$ ,  $\sigma = \$3.61$

b)  $\mu = \$4.00$ ,  $\sigma = \$5.10$

c) 0.191. Model is  $N(80, 22.83)$ .

43. a)  $\mu = 2.859$ ,  $\sigma = 1.324$

b) No. The score distribution in the sample should resemble that in the population, somewhat uniform for scores 1–4 and about half as many 5's.

c) Approximately  $N\left(2.859, \frac{1.324}{\sqrt{40}}\right)$ .

45. About 20%, based on  $N(2.859, 0.167)$ .

47. a)  $N(2.9, 0.045)$       b) 0.0131      c) 2.97 gm/mi

49. a) Can't use a Normal model to estimate probabilities. The distribution is skewed right—not Normal.

b) 4 is probably not a large enough sample to say the average follows the Normal model.

c) No. This is 3.16 SDs above the mean.

51. a) 0.0003. Model is  $N(384, 34.15)$ .      b) \$427.77 or more.

53. a) 0.734

b) 0.652. Model is  $N(10, 12.81)$ .

c) 0.193. Model is  $N(120, 5.774)$ .

d) 0.751. Model is  $N(10, 7.394)$ .

## CHAPTER 19

1. She believes the true proportion is within 4% of her estimate, with some (probably 95%) degree of confidence.

3. a) Population—all cars; sample—those actually stopped at the checkpoint;  $p$ —proportion of all cars with safety problems;

- $\hat{p}$ —proportion actually seen with safety problems (10.4%); if sample (a cluster sample) is representative, then the methods of this chapter will apply.
- b) Population—general public; sample—those who logged onto the Web site;  $p$ —population proportion of those who favor prayer in school;  $\hat{p}$ —proportion of those who voted in the poll who favored prayer in school (81.1%); can't use methods of this chapter—sample is biased and nonrandom.
- c) Population—parents at the school; sample—those who returned the questionnaire;  $p$ —proportion of all parents who favor uniforms;  $\hat{p}$ —proportion of respondents who favor uniforms (60%); should not use methods of this chapter, since not SRS (possible non-response bias).
- d) Population—students at the college; sample—the 1632 students who entered that year;  $p$ —proportion of all students who will graduate on time;  $\hat{p}$ —proportion of that year's students who graduate on time (85.0%); can use methods of this chapter if that year's students (a cluster sample) are viewed as a representative sample of all possible students at the school.
5. a) Not correct. This implies certainty.  
 b) Not correct. Different samples will give different results. Many fewer than 95% will have 88% on-time orders.  
 c) Not correct. The interval is about the population proportion, not the sample proportion in different samples.  
 d) Not correct. In this sample, we *know* 88% arrived on time.  
 e) Not correct. The interval is about the parameter, not the days.
7. a) False    b) True    c) True    d) False
9. On the basis of this sample, we are 90% confident that the proportion of Japanese cars is between 29.9% and 47.0%.
11. a) (0.798, 0.863)  
 b) We're 95% confident that between 80% and 86% of all broiler chicken sold in U.S. food stores is infected with *Campylobacter*.  
 c) The size of the population is irrelevant. If *Consumer Reports* had a random sample, 95% of intervals generated by studies like this will capture the true contamination level.
13. a) 0.025  
 b) We're 90% confident that this poll's estimate is within  $\pm 2.5\%$  of the true proportion of people who are baseball fans.  
 c) Larger. To be more certain, we must be less precise.  
 d) 0.039    e) less confidence  
 f) No evidence of change; given the margin of error, 0.37 is a plausible value for 2007 as well.
15. a) (0.0465, 0.0491). The assumptions and conditions for constructing a confidence interval are satisfied.  
 b) The confidence interval gives the set of plausible values (with 95% confidence). Since 0.05 is outside the interval, that seems to be a bit too optimistic.
17. a) (12.7%, 18.6%)  
 b) We are 95% confident, based on this sample, that the proportion of all auto accidents that involve teenage drivers is between 12.7% and 18.6%.  
 c) About 95% of all random samples will produce confidence intervals that contain the true population proportion.  
 d) Contradicts. The interval is completely below 20%.
19. Probably nothing. Those who bothered to fill out the survey may be a biased sample.
21. a) Response bias (wording)    b) (54%, 60%)  
 c) Smaller—the sample size was larger.
23. a) (18.2%, 21.8%)  
 b) We are 98% confident, based on the sample, that between 18.2% and 21.8% of English children are deficient in vitamin D.  
 c) About 98% of all random samples will produce a confidence interval that contains the true proportion of children deficient in vitamin D.
25. a) Wider. The sample size is probably about one-fourth of the sample size for all adults, so we'd expect the confidence interval to be about twice as wide.  
 b) Smaller. The second poll used a slightly larger sample size.

27. a) (15.5%, 26.3%)    b) 612  
 c) Sample may not be random or representative. Deer that are legally hunted may not represent all sexes and ages.
29. a) 141    b) 318    c) 564
31. 1801    33. 384 total, using  $p = 0.15$     35. 90%

## CHAPTER 20

1. a)  $H_0: p = 0.30; H_A: p < 0.30$   
 b)  $H_0: p = 0.50; H_A: p \neq 0.50$   
 c)  $H_0: p = 0.20; H_A: p > 0.20$
3. Statement d is correct.
5. No, we can say only that there is a 27% chance of seeing the observed effectiveness just from natural sampling variation. There is no *evidence* that the new formula is more effective, but we can't conclude that they are equally effective.
7. a) No. There's a 25% chance of losing twice in a row. That's not unusual.  
 b) 0.125    c) No, we expect that to happen 1 time in 8.  
 d) Maybe 5? The chance of 5 losses in a row is only 1 in 32, which seems unusual.
9. 1) Use  $p$ , not  $\hat{p}$ , in hypotheses.  
 2) The question was about failing to meet the goal, so  $H_A$  should be  $p < 0.96$ .  
 3) Did not check  $0.04(200) = 8$ . Since  $nq < 10$ , the Success/Failure Condition is violated. Didn't check 10% Condition.
- 4)  $188/200 = 0.94; SD(\hat{p}) = \sqrt{\frac{(0.96)(0.04)}{200}} = 0.014$
- 5)  $z$  is incorrect; should be  $z = \frac{0.94 - 0.96}{0.014} = -1.43$
- 6)  $P = P(z < -1.43) = 0.076$
- 7) There is only weak evidence that the new instructions do not work.
11. a)  $H_0: p = 0.30; H_A: p > 0.30$   
 b) Possibly an SRS; we don't know if the sample is less than 10% of his customers, but it could be viewed as less than 10% of all possible customers;  $(0.3)(80) \geq 10$  and  $(0.7)(80) \geq 10$ . Wells are independent only if customers don't have farms on the same underground springs.  
 c)  $z = 0.73; P\text{-value} = 0.232$   
 d) If his dowsing is no different from standard methods, there is more than a 23% chance of seeing results as good as those of the dowser's, or better, by natural sampling variation.  
 e) These data provide no evidence that the dowser's chance of finding water is any better than normal drilling.
13. a)  $H_0: p_{2000} = 0.34; H_A: p_{2000} \neq 0.34$   
 b) Students were randomly sampled and should be independent. 34% and 66% of 8302 are greater than 10. 8302 students is less than 10% of the entire student population of the United States.  
 c)  $P = 0.058$   
 d) With such a small P-value, I reject  $H_0$ . There has been a statistically significant change in the proportion of students who have no absences.  
 e) No. A difference this small, although statistically significant, is not meaningful. We might look at new data in a few years.
15. a)  $H_0: p = 0.05$  vs.  $H_A: p < 0.05$   
 b) We assume the whole mailing list has over 1,000,000 names. This is a random sample, and we expect 5000 successes and 95,000 failures.  
 c)  $z = -3.178; P\text{-value} = 0.00074$ , so we reject  $H_0$ ; there is strong evidence that the donation rate would be below 5%.
17. a)  $H_0: p = 0.63; H_A: p > 0.63$   
 b) The sample is representative.  $240 < 10\%$  of all law school applicants. We expect  $240(0.63) = 151.2$  to be admitted and  $240(0.37) = 88.8$  not to be, both at least 10.  $z = 1.58; P\text{-value} = 0.057$

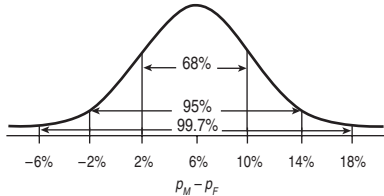
- c) Although the evidence is weak, there is some indication that the program may be successful. Candidates should decide whether they can afford the time and expense.
19.  $H_0: p = 0.20$ ;  $H_A: p > 0.20$ . SRS (not clear from information provided); 22 is more than 10% of the population of 150;  $(0.20)(22) < 10$ . Do not proceed with a test.
21.  $H_0: p = 0.03$ ;  $p \neq 0.03$ .  $\hat{p} = 0.015$ . One mother having twins will not affect another, so observations are independent; not an SRS; sample is less than 10% of all births. However, the mothers at this hospital may not be representative of all teenagers;  $(0.03)(469) = 14.07 \geq 10$ ;  $(0.97)(469) \geq 10$ .  $z = -1.91$ ; P-value = 0.0556. With a P-value this low, reject  $H_0$ . These data show some evidence that the rate of twins born to teenage girls at this hospital is less than the national rate of 3%. It is not clear whether this can be generalized to all teenagers.
23.  $H_0: p = 0.25$ ;  $H_A: p > 0.25$ . SRS; sample is less than 10% of all potential subscribers;  $(0.25)(500) \geq 10$ ;  $(0.75)(500) \geq 10$ .  $z = 1.24$ ; P-value = 0.1076. The P-value is high, so do not reject  $H_0$ . These data do not show that more than 25% of current readers would subscribe; the company should not go ahead with the WebZine on the basis of these data.
25.  $H_0: p = 0.40$ ;  $H_A: p < 0.40$ . Data are for all executives in this company and may not be able to be generalized to all companies;  $(0.40)(43) \geq 10$ ;  $(0.60)(43) \geq 10$ .  $z = -1.31$ ; P-value = 0.0955. Because the P-value is high, we fail to reject  $H_0$ . These data do not show that the proportion of women executives is less than the 40% of women in the company in general.
27.  $H_0: p = 0.103$ ;  $H_A: p > 0.103$ .  $\hat{p} = 0.118$ ;  $z = 2.06$ ; P-value = 0.02. Because the P-value is low, we reject  $H_0$ . These data provide evidence that the dropout rate has increased.
29.  $H_0: p = 0.90$ ;  $H_A: p < 0.90$ .  $\hat{p} = 0.844$ ;  $z = -2.05$ ; P-value = 0.0201. Because the P-value is so low, we reject  $H_0$ . There is strong evidence that the actual rate at which passengers with lost luggage are reunited with it within 24 hours is less than the 90% claimed by the airline.
31. a) Yes; assuming this sample to be a typical group of people,  $P = 0.0008$ . This cancer rate is very unusual.  
b) No, this group of people may be atypical for reasons that have nothing to do with the radiation.
9. a) (1.9%, 4.1%)  
b) Because 5% is not in the interval, there is strong evidence that fewer than 5% of all men use work as their primary measure of success.  
c)  $\alpha = 0.01$ ; it's a lower-tail test based on a 98% confidence interval.
11. a) (0.274, 0.327)  
b) Since 0.27 is not in the confidence interval, we reject the hypothesis that  $p = 0.27$
13. a) The Success/Failure Condition is violated: only 5 pups had dysplasia.  
b) We are 95% confident that between 5% and 26% of puppies will show signs of hip dysplasia at the age of 6 months.
15. a) Type II error    b) Type I error  
c) By making it easier to get the loan, the bank has reduced the alpha level.  
d) The risk of a Type I error is decreased and the risk of a Type II error is increased.
17. a) Power is the probability that the bank denies a loan that would not have been repaid.  
b) Raise the cutoff score.  
c) A larger number of trustworthy people would be denied credit, and the bank would miss the opportunity to collect interest on those loans.
19. a) The null is that the level of home ownership remains the same. The alternative is that it rises.  
b) The city concludes that home ownership is on the rise, but in fact the tax breaks don't help.  
c) The city abandons the tax breaks, but they were helping.  
d) A Type I error causes the city to forego tax revenue, while a Type II error withdraws help from those who might have otherwise been able to buy a home.  
e) The power of the test is the city's ability to detect an actual increase in home ownership.
21. a) It is decided that the shop is not meeting standards when it is.  
b) The shop is certified as meeting standards when it is not.  
c) Type I    d) Type II
23. a) The probability of detecting a shop that is not meeting standards.  
b) 40 cars. Larger  $n$ .    c) 10%. More chance to reject  $H_0$ .  
d) A lot. Larger differences are easier to detect.
25. a) One-tailed. The company wouldn't be sued if "too many" minorities were hired.  
b) Deciding the company is discriminating when it is not.  
c) Deciding the company is not discriminating when it is.  
d) The probability of correctly detecting actual discrimination.  
e) Increases power.    f) Lower, since  $n$  is smaller.
27. a) One-tailed. Software is supposed to decrease the dropout rate.  
b)  $H_0: p = 0.13$ ;  $H_A: p < 0.13$   
c) He buys the software when it doesn't help students.  
d) He doesn't buy the software when it does help students.  
e) The probability of correctly deciding the software is helpful.
29. a)  $z = -3.21$ ,  $p = 0.0007$ . The change is statistically significant. A 95% confidence interval is (2.3%, 8.5%). This is clearly lower than 13%. If the cost of the software justifies it, the professor should consider buying the software.  
b) The chance of observing 11 or fewer dropouts in a class of 203 is only 0.07% if the dropout rate is really 13%.
31. a)  $H_A: p = 0.30$ , where  $p$  is the probability of heads  
b) Reject the null hypothesis if the coin comes up tails—otherwise fail to reject.  
c)  $P(\text{tails given the null hypothesis}) = 0.1 = \alpha$   
d)  $P(\text{tails given the alternative hypothesis}) = \text{power} = 0.70$   
e) Spin the coin more than once and base the decision on the sample proportion of heads.
33. a) 0.0464    b) Type I    c) 37.6%  
d) Increase the number of shots. Or keep the number of shots at 10, but increase alpha by declaring that 8, 9, or 10 will be deemed as having improved.

## CHAPTER 21

1. a) Two sided. Let  $p$  be the percentage of students who prefer Diet Pepsi.  $H_0: p = 0.5$  vs.  $H_A: p \neq 0.5$   
b) One sided. Let  $p$  be the percentage of teenagers who prefer the new formulation.  $H_0: p = 0.5$  vs.  $H_A: p > 0.5$   
c) One sided. Let  $p$  be the percentage of people who intend to vote for the override.  $H_0: p = 2/3$  vs.  $H_A: p > 2/3$ .  
d) Two sided. Let  $p$  be the percentage of days that the market goes up.  $H_0: p = 0.5$  vs.  $H_A: p \neq 0.5$
3. If there is no difference in effectiveness, the chance of seeing an observed difference this large or larger is 4.7% by natural sampling variation.
5.  $\alpha = 0.10$ : Yes. The P-value is less than 0.05, so it's less than 0.10. But to reject  $H_0$  at  $\alpha = 0.01$ , the P-value must be below 0.01, which isn't necessarily the case.
7. a) There is only a 1.1% chance of seeing a sample proportion as low as 89.4% vaccinated by natural sampling variation if 90% have really been vaccinated.  
b) We conclude that  $p$  is below 0.9, but a 95% confidence interval would suggest that the true proportion is between (0.889, 0.899). Most likely, a decrease from 90% to 89.9% would not be considered important. On the other hand, with 1,000,000 children a year vaccinated, even 0.1% represents about 1000 kids—so this may very well be important.

## CHAPTER 22

- It's very unlikely that samples would show an observed difference this large if in fact there is no real difference in the proportions of boys and girls who have used online social networks.
- The ads may be working. If there had been no real change in name recognition, there'd be only about a 3% chance the percentage of voters who heard of this candidate would be at least this much higher in a different sample.
- The responses are not from two independent groups, but are from the same individuals.
- a) Stratified    b) 6% higher among males    c) 4%  
d)



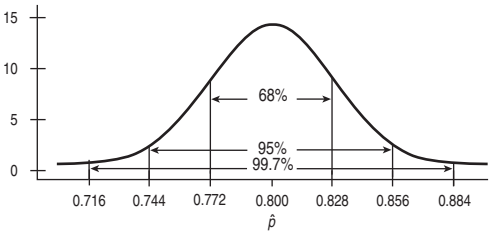
- Yes; a poll result showing little difference is only 1–2 standard deviations below the expected outcome.
- a) Yes. Random sample; less than 10% of the population; samples are independent; more than 10 successes and failures in each sample.  
b) (0.055, 0.140)  
c) We are 95% confident, based on these samples, that the proportion of American women age 65 and older who suffer from arthritis is between 5.5% and 14.0% more than the proportion of American men of the same age who suffer from arthritis.  
d) Yes; the entire interval lies above 0.
- a) 0.035    b) (0.356, 0.495)  
c) We are 95% confident, based on these data, that the proportion of pets with a malignant lymphoma in homes where herbicides are used is between 35.6% and 49.5% higher than the proportion of pets with lymphoma in homes where no pesticides are used.
- a) Yes, subjects were randomly divided into independent groups, and more than 10 successes and failures were observed in each group.  
b) (4.7%, 8.9%)  
c) Yes, we're 95% confident that the rate of infection is 5–9 percentage points lower. That's a meaningful reduction, considering the 20% infection rate among the unvaccinated kids.
- a)  $H_0: p_V - p_{NV} = 0$ ,  $H_A: p_V - p_{NV} < 0$ .  
b) Because 0 is not in the confidence interval, reject the null. There's evidence that the vaccine reduces the rate of ear infections.  
c) 2.5%    d) Type I  
e) Babies would be given ineffective vaccinations.
- a) Prospective study  
b)  $H_0: p_1 - p_2 = 0$ ;  $H_A: p_1 - p_2 \neq 0$  where  $p_1$  is the proportion of students whose parents disapproved of smoking who became smokers and  $p_2$  is the proportion of students whose parents are lenient about smoking who became smokers.  
c) Yes. We assume the students were randomly selected; they are less than 10% of the population; samples are independent; at least 10 successes and failures in each sample.  
d)  $z = -1.17$ , P-value = 0.2422. These samples do not show evidence that parental attitudes influence teens' decisions to smoke.  
e) If there is no difference in the proportions, there is about a 24% chance of seeing the observed difference or larger by natural sampling variation.  
f) Type II
- a) (-0.065, 0.221)  
b) We are 95% confident that the proportion of teens whose parents disapprove of smoking who will eventually smoke is between 22.1% less and 6.5% more than for teens with parents who are lenient about smoking.

- 95% of all random samples will produce intervals that contain the true difference.
- a) No; subjects weren't assigned to treatment groups. It's an observational study.  
b)  $H_0: p_1 - p_2 = 0$ ;  $H_A: p_1 - p_2 \neq 0$ .  $z = 3.56$ , P-value = 0.0004. With a P-value this low, we reject  $H_0$ . There is a significant difference in the clinic's effectiveness. Younger mothers have a higher birth rate than older mothers. Note that the Success/Failure Condition is met based on the pooled estimate of  $p$ .  
c) We are 95% confident, based on these data, that the proportion of successful live births at the clinic is between 10.0% and 27.8% higher for mothers under 38 than in those 38 and older. However, the Success/Failure Condition is not met for the older women, since # Successes < 10. We should be cautious in trusting this confidence interval.
- a)  $H_0: p_1 - p_2 = 0$ ;  $H_A: p_1 - p_2 > 0$ .  $z = 1.18$ , P-value = 0.118. With P-value this high, we fail to reject  $H_0$ . These data do not show evidence of a decrease in the voter support for the candidate.  
b) Type II
- a)  $H_0: p_1 - p_2 = 0$ ;  $H_A: p_1 - p_2 \neq 0$ .  $z = -0.39$ , P-value = 0.6951. With a P-value this high, we fail to reject  $H_0$ . There is no evidence of racial differences in the likelihood of multiple births, based on these data.  
b) Type II
- a) We are 95% confident, that between 67.0% and 83.0% of patients with joint pain will find medication A effective.  
b) We are 95% confident, that between 51.9% and 70.3% of patients with joint pain will find medication B effective.  
c) Yes, they overlap. This might indicate no difference in the effectiveness of the medications. (Not a proper test.)  
d) We are 95% confident that the proportion of patients with joint pain who will find medication A effective is between 1.7% and 26.1% higher than the proportion who will find medication B effective.  
e) No. There is a difference in the effectiveness of the medications.  
f) To estimate the variability in the difference of proportions, we must add variances. The two one-sample intervals do not. The two-sample method is the correct approach.
- The conditions are satisfied to test  $H_0: p_{\text{young}} = p_{\text{old}}$  against  $H_A: p_{\text{young}} > p_{\text{old}}$ . The one-sided P-value is 0.0619, so we may reject the null hypothesis. Although the evidence is not strong, *Time* may be justified in saying that younger men are more comfortable discussing personal problems.
- Yes. With a low P-value of 0.003, reject the null hypothesis of no difference. There's evidence of an increase in the proportion of parents checking the Web sites visited by their teens.

## PART V REVIEW

- $H_0$ : There is no difference in cancer rates,  $p_1 - p_2 = 0$ .  $H_A$ : The cancer rate in those who use the herb is higher,  $p_1 - p_2 > 0$ .
- a) 10.29  
b) Not really. The z-score is -1.11. Not any evidence to suggest that the proportion for Monday is low.  
c) Yes. The z-score is 2.26 with a P-value of 0.024 (two-sided).  
d) Some births are scheduled for the convenience of the doctor and/or the mother.
- a)  $H_0: p_1 = 0.40$ ;  $H_A: p_1 < 0.40$   
b) Random sample; less than 10% of all California gas stations,  $0.4(27) = 10.8$ ,  $0.6(27) = 16.2$ . Assumptions and conditions are met.  
c)  $z = -1.49$ , P-value = 0.0677  
d) With a P-value this high, we fail to reject  $H_0$ . These data do not provide evidence that the proportion of leaking gas tanks is less than 40% (or that the new program is effective in decreasing the proportion).



- e) Yes, Type II.
  - f) Increase  $\alpha$ , increase the sample size.
  - g) Increasing  $\alpha$ —increases power, lowers chance of Type II error, but increases chance of Type I error. Increasing sample size—increases power, costs more time and money.
7. a) The researcher believes that the true proportion of “A’s” is within 10% of the estimated 54%, namely, between 44% and 64%.  
b) Small sample    c) No, 63% is contained in the interval.
  9. a) Pew believes that the true proportion is within 3% of the 33% from the sample; that is, between 30% and 36%.  
b) Larger, since it’s a smaller sample.  
c) We are 95% confident that the proportion of active traders who rely on the Internet for investment information is between 38.7% and 51.3%, based on this sample.  
d) Larger, since it’s a smaller sample.
  11. a) Bimodal!  
b)  $\mu$ , the population mean. Sample size does not matter.  
c)  $\sigma/\sqrt{n}$ ; sample size does matter.  
d) It becomes closer to a Normal model and narrower as the sample size increases.
  13. a)  $\mu = 0.80, \sigma = 0.028$   
b) Yes.  $0.8(200) = 160, 0.2(200) = 40$ . Both  $\geq 10$ .  
c)   
d) 0.039
  15.  $H_0$ : There is no difference,  $p_1 - p_2 = 0$ .  $H_A$ : Early births have increased,  $p_1 - p_2 < 0$ .  $z = -0.729$ , P-value = 0.2329. Because the P-value is so high, we do not reject  $H_0$ . These data do not show an increase in the incidence of early birth of twins.
  17. a)  $H_0$ : There is no difference,  $p_1 - p_2 \geq 0$ .  $H_A$ : Treatment prevents deaths from eclampsia,  $p_1 - p_2 < 0$ .  
b) Samples are random and independent; less than 10% of all pregnancies (or eclampsia cases); more than 10 successes and failures in each group.  
c) 0.8008  
d) There is insufficient evidence to conclude that magnesium sulfide is effective in preventing eclampsia deaths.  
e) Type II    f) Increase the sample size, increase  $\alpha$ .  
g) Increasing sample size: decreases variation in the sampling distribution, is costly. Increasing  $\alpha$ : Increases likelihood of rejecting  $H_0$ , increases chance of Type I error.
  19. a) It is not clear what the pollster asked. Otherwise they did fine.  
b) Stratified sampling.    c) 4%  
d) 95%    e) Smaller sample size.  
f) Wording and order of questions (response bias).
  21. a)  $H_0$ : There is no difference,  $p = 0.143$ .  $H_A$ : The fatal accident rate is lower in girls,  $p < 0.143$ .  $z = -1.67$ , P-value = 0.0479. Because the P-value is low, we reject  $H_0$ . These data give some evidence that the fatal accident rate is lower for girls than for teens in general.  
b) If the proportion is really 14.3%, we will see the observed proportion (11.3%) or lower 4.8% of the time by sampling variation.
  23. a) One would expect many small fish, with a few large ones.  
b) We don’t know the exact distribution, but we know it’s not Normal.  
c) Probably not. With a skewed distribution, a sample size of five is not a large enough sample to say the sampling model for the mean is approximately Normal.  
d) 0.961

25. a) Yes.  $0.8(60) = 48, 0.2(60) = 12$ . Both are  $\geq 10$ .  
b) 0.834  
c) Higher. Bigger sample means smaller standard deviation for  $\hat{p}$ .  
d) Answers will vary. For  $n = 500$ , the probability is 0.997.
27. a) 54.4 to 62.5%  
b) Based on this study, with 95% confidence the proportion of Crohn’s disease patients who will respond favorable to infliximab is between 54.4% and 62.5%.  
c) 95% of all such random samples will produce confidence intervals that contain the true proportion of patients who respond favorably.
29. At least 423, assuming that  $p$  is near 50%.
31. a) Random sample (?); certainly less than 10% of all preemies and normal babies; more than 10 failures and successes in each group. 1.7% to 16.3% greater for normal-birth weight children.  
b) Since 0 is not in the interval, there is evidence that preemies have a lower high school graduation rate than children of normal birth weight.  
c) Type I, since we rejected the null hypothesis.
33. a)  $H_0$ : The computer is undamaged.  $H_A$ : The computer is damaged.  
b) 20% of good PCs will be classified as damaged (bad), while all damaged PCs will be detected (good).  
c) 3 or more.    d) 20%  
e) By switching to two or more as the rejection criterion, 7% of the good PCs will be misclassified, but only 10% of the bad ones will, increasing the power from 20% to 90%.
35. The null hypothesis is that Bush’s disapproval proportion is 66%—the Nixon benchmark. The one-tailed test has a z-value of  $-2.00$ , so the P-value is 0.0228. It looks like Bush’s May 2007 ratings were better than the Nixon benchmark low.
37. a) The company is interested only in confirming that the athlete is well known.  
b) Type I: the company concludes that the athlete is well known, but that’s not true. It offers an endorsement contract to someone who lacks name recognition. Type II: the company overlooks a well-known athlete, missing the opportunity to sign a potentially effective spokesperson.  
c) Type I would be more likely, Type II less likely.
39. I am 95% confident that the proportion of U.S. adults who favor nuclear energy is between 7 and 19 percentage points higher than the proportion who would accept a nuclear plant near their area.

**CHAPTER 23**

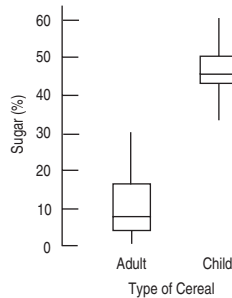
1. a) 1.74    b) 2.37    c) 0.0524    d) 0.0889
3. Shape becomes closer to Normal; center does not change; spread becomes narrower.
5. a) The confidence interval is for the population mean, not the individual cows in the study.  
b) The confidence interval is not for individual cows.  
c) We *know* the average gain in this study was 56 pounds!  
d) The average weight gain of all cows does not vary. It’s what we’re trying to estimate.  
e) No. There is not a 95% chance for another sample to have an average weight gain between 45 and 67 pounds. There is a 95% chance that another sample will have its average weight gain within two standard errors of the true mean.
7. a) No. A confidence interval is not about individuals in the population.  
b) No. It’s not about individuals in the sample, either.  
c) No. We know the mean cost for students in the sample was \$1196.  
d) No. A confidence interval is not about other sample means.  
e) Yes. A confidence interval estimates a population parameter.

9. a) Based on this sample, we can say, with 95% confidence, that the mean pulse rate of adults is between 70.9 and 74.5 beats per minute.  
 b) 1.8 beats per minute  
 c) Larger
11. The assumptions and conditions for a  $t$ -interval are not met. The distribution is highly skewed to the right and there is a large outlier.
13. a) Yes. Randomly selected group; less than 10% of the population; the histogram is not unimodal and symmetric, but it is not highly skewed and there are no outliers, so with a sample size of 52, the CLT says  $\bar{y}$  is approximately Normal.  
 b) (98.06, 98.51) degrees F  
 c) We are 98% confident, based on the data, that the average body temperature for an adult is between 98.06°F and 98.51°F.  
 d) 98% of all such random samples will produce intervals containing the true mean temperature.  
 e) These data suggest that the true normal temperature is somewhat less than 98.6°F.
15. a) Narrower. A smaller margin of error, so less confident.  
 b) Advantage: more chance of including the true value. Disadvantage: wider interval.  
 c) Narrower; due to the larger sample, the SE will be smaller.  
 d) About 252
17. a) (709.90, 802.54)  
 b) With 95% confidence, based on these data, the speed of light is between 299,709.9 and 299,802.5 km/sec.  
 c) Normal model for the distribution, independent measurements. These seem reasonable here, but it would be nice to see if the Nearly Normal Condition held for the data.
19. a) Given no time trend, the monthly on-time departure rates should be independent. Though not a random sample, these months should be representative, and they're fewer than 10% of all months. The histogram looks unimodal, but slightly left-skewed; not a concern with this large sample.  
 b)  $80.57 < \mu(\text{OT Departure}\%) < 81.80$   
 c) We can be 90% confident that the interval from 80.57% to 81.80% holds the true mean monthly percentage of on-time flight departures.
21. The 95% confidence interval lies entirely above the 0.08 ppm limit, evidence that mirex contamination is too high and consistent with rejecting the null. We used an upper-tail test, so the P-value should therefore be smaller than  $\frac{1}{2}(1 - 0.95) = 0.025$ , and it was.
23. If in fact the mean cholesterol of pizza eaters does not indicate a health risk, then only 7 of every 100 samples would have mean cholesterol levels as high (or higher) as observed in this sample.
25. a) Upper-tail. We want to show it will hold 500 pounds (or more) easily.  
 b) They will decide the stands are safe when they're not.  
 c) They will decide the stands are unsafe when they are in fact safe.
27. a) Decrease  $\alpha$ . This means a smaller chance of declaring the stands safe if they are not.  
 b) The probability of correctly detecting that the stands are capable of holding more than 500 pounds.  
 c) Decrease the standard deviation—probably costly. Increase the sample size—takes more time for testing and is costly. Increase  $\alpha$ —more Type I errors. Increase the “design load” to be well above 500 pounds—again, costly.
29. a)  $H_0: \mu = 23.3; H_A: \mu > 23.3$   
 b) We have a random sample of the population. Population may not be normally distributed, as it would be easier to have a few much older men at their first marriage than some very young men. However, with a sample size of 40,  $\bar{y}$  should be approximately Normal. We should check the histogram for severity of skewness and possible outliers.
- c)  $(\bar{y} - 23.3)/(s/\sqrt{40}) \sim t_{39}$  d) 0.1447
- e) If the average age at first marriage is still 23.3 years, there is a 14.5% chance of getting a sample mean of 24.2 years or older simply from natural sampling variation.
- f) We lack evidence that the average age at first marriage has increased from the mean of 23.3 years.
31. a) Probably a representative sample; the Nearly Normal Condition seems reasonable. (Show a Normal probability plot or histogram.) The histogram is nearly uniform, with no outliers or skewness.  
 b)  $\bar{y} = 28.78, s = 0.40$  c) (28.36, 29.21) grams  
 d) Based on this sample, we are 95% confident the average weight of the content of Ruffles bags is between 28.36 and 29.21 grams.  
 e) The company is erring on the safe side, as it appears that, on average, it is putting in slightly more chips than stated.
33. a) Type I; he mistakenly rejected the null hypothesis that  $p = 0.10$  (or worse).  
 b) Yes. These are a random sample of bags and the Nearly Normal Condition is met (Show a Normal probability plot or histogram.);  $t = -2.51$  with 7 df for a one-sided P-value of 0.0203.
35. a) Random sample; the Nearly Normal Condition seems reasonable from a Normal probability plot. The histogram is roughly unimodal and symmetric with no outliers. (Show plot.)  
 b) (1187.9, 1288.4) chips  
 c) Based on this sample, the mean number of chips in an 18-ounce bag is between 1187.9 and 1288.4, with 95% confidence. The *mean* number of chips is clearly greater than 1000. However, if the claim is about individual bags, then it's not necessarily true. If the mean is 1188 and the SD deviation is near 94, then 2.5% of the bags will have fewer than 1000 chips, using the Normal model. If in fact the mean is 1288, the proportion below 1000 will be less than 0.1%, but the claim is still false.
37. a) The Normal probability plot is relatively straight, with one outlier at 93.8 sec. Without the outlier, the conditions seem to be met. The histogram is roughly unimodal and symmetric with no other outliers. (Show your plot.)  
 b)  $t = -2.63$ , P-value = 0.0160. With the outlier included, we might conclude that the mean completion time for the maze is not 60 seconds; in fact, it is less.  
 c)  $t = -4.46$ , P-value = 0.0003. Because the P-value is so small, we reject  $H_0$ . Without the outlier, we see strong evidence that the average completion time for the maze is less than 60 seconds. The outlier here did not change the conclusion.  
 d) The maze does not meet the “one-minute average” requirement. Both tests rejected a null hypothesis of a mean of 60 seconds.
39. a)  $287.3 < \mu(\text{Drive Distance}) < 289.9$   
 b) These data are not a random sample of golfers. The top professionals are (unfortunately) not representative and were not selected at random. We might consider the 2006 data to represent the population of all professional golfers, past, present, and future.  
 c) The data are means for each golfer, so they are less variable than if we looked at all the separate drives.

## CHAPTER 24

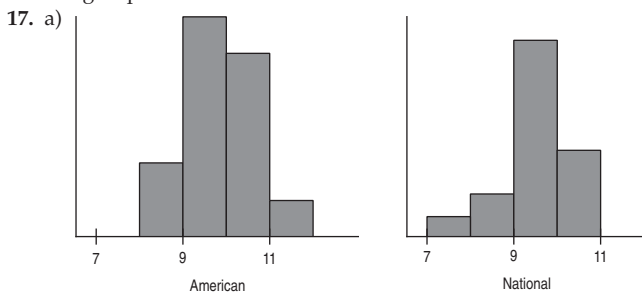
1. Yes. The high P-value means that we lack evidence of a difference, so 0 is a possible value for  $\mu_{\text{Meat}} - \mu_{\text{Beef}}$ .
3. a) Plausible values of  $\mu_{\text{Meat}} - \mu_{\text{Beef}}$  are all negative, so the mean fat content is probably higher for beef hot dogs.  
 b) The difference is significant. c) 10%
5. a) False. The confidence interval is about means, not about individual hot dogs.  
 b) False. The confidence interval is about means, not about individual hot dogs.

- c) True.
- d) False. CI's based on other samples will also try to estimate the true difference in population means; there's no reason to expect other samples to conform to this result.
- e) True.
- 7. a) 2.927    b) Larger
- c) Based on this sample, we are 95% confident that students who learn Math using the CPMP method will score, on average, between 5.57 and 11.43 points better on a test solving applied Algebra problems with a calculator than students who learn by traditional methods.
- d) Yes; 0 is not in the interval.
- 9. a)  $H_0: \mu_C - \mu_T = 0$  vs.  $H_A: \mu_C - \mu_T \neq 0$
- b) Yes. Groups are independent, though we don't know if students were randomly assigned to the programs. Sample sizes are large, so CLT applies.
- c) If the means for the two programs are really equal, there is less than a 1 in 10,000 chance of seeing a difference as large as or larger than the observed difference just from natural sampling variation.
- d) On average, students who learn with the CPMP method do significantly worse on Algebra tests that do not allow them to use calculators than students who learn by traditional methods.
- 11. a) (1.36, 4.64)
- b) No; 5 minutes is beyond the high end of the interval.
- 13.



Random sample—questionable, but probably representative, independent samples, less than 10% of all cereals; boxplot shows no outliers—not exactly symmetric, but these are reasonable sample sizes. Based on these samples, with 95% confidence, children's cereals average between 32.49% and 40.80% more sugar content than adult's cereals.

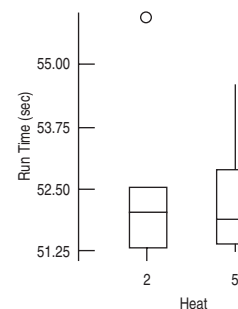
- 15.  $H_0: \mu_N - \mu_C = 0$  vs.  $H_A: \mu_N - \mu_C > 0$ ;  $t = 2.207$ ; P-value = 0.0168; df = 33.4. Because of the small P-value, we reject  $H_0$ . These data do suggest that new activities are better. The mean reading comprehension score for the group with new activities is significantly (at  $\alpha = 0.05$ ) higher than the mean score for the control group.



Both are unimodal and reasonably symmetric.

- b) Based on these data, the average number of runs in an American League stadium is between 9.36 and 10.23, with 95% confidence.
- c) No. The boxplot indicates it isn't an outlier.

- d) We want to work directly with the average difference. The two separate confidence intervals do not answer questions about the difference. The difference has a different standard deviation, found by adding variances.
- 19. a) (-0.18, 0.89)
- b) Based on these data, with 95% confidence, American League stadiums average between 0.18 fewer runs and 0.89 more runs per game than National League stadiums.
- c) No; 0 is in the interval.
- 21. These are not two independent samples. These are before and after scores for the same individuals.
- 23. a) These data do not provide evidence of a difference in ad recall between shows with sexual content and violent content.
- b)  $H_0: \mu_S - \mu_N = 0$  vs.  $H_A: \mu_S - \mu_N \neq 0$ .  $t = -6.08$ , df = 213.99, P-value =  $5.5 \times 10^{-9}$ . Because the P-value is low, we reject  $H_0$ . These data suggest that ad recall between shows with sexual and neutral content is different; those who saw shows with neutral content had higher average recall.
- 25. a)  $H_0: \mu_V - \mu_N = 0$  vs.  $H_A: \mu_V - \mu_N \neq 0$ .  $t = -7.21$ , df = 201.96, P-value =  $1.1 \times 10^{-11}$ . Because of the very small P-value, we reject  $H_0$ . There is a significant difference in mean ad recall between shows with violent content and neutral content; viewers of shows with neutral content remember more brand names, on average.
- b) With 95% confidence, the average number of brand names remembered 24 hours later is between 1.45 and 2.41 higher for viewers of neutral content shows than for viewers of sexual content shows, based on these data.
- 27.  $H_0: \mu_{big} - \mu_{small} = 0$  vs.  $H_A: \mu_{big} - \mu_{small} \neq 0$ ; bowl size was assigned randomly; amount scooped by individuals and by the two groups should be independent. With 34.3 df,  $t = 2.104$  and P-value = 0.0428. The low P-value leads us to reject the null hypothesis. There is evidence of a difference in the average amount of ice cream that people scoop when given a bigger bowl.
- 29. a) The 95% confidence interval for the difference is (0.61, 5.39). 0 is not in the interval, so scores in 1996 were significantly higher. (Or the  $t$ , with more than 7500 df, is 2.459 for a P-value of 0.0070.)
- b) Since both samples were very large, there shouldn't be a difference in how certain you are, assuming conditions are met.
- 31. Independent Groups Assumption: The runners are different women, so the groups are independent. The Randomization Condition is satisfied since the runners are selected at random for these heats.



Nearly Normal Condition: The boxplots show an outlier, but we will proceed and then redo the analysis with the outlier deleted. When we include the outlier,  $t = 0.035$  with a two-sided P-value of 0.97. With the outlier deleted,  $t = -1.14$ , with  $P = 0.2837$ .

Either P-value is so large that we fail to reject the null hypothesis of equal means and conclude that there is no evidence of a difference in the mean times for runners in unseeded heats.

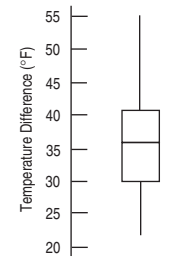
- 33. With  $t = -4.57$  and a very low P-value of 0.0013, we reject the null hypothesis of equal mean velocities. There is strong evidence that golf balls hit off Stinger tees will have a higher mean initial velocity.

35. a) We can be 95% confident that the interval  $74.8 \pm 178.05$  minutes includes the true difference in mean crossing times between men and women. Because the interval includes zero, we cannot be confident that there is any difference at all.
- b) Independence Assumption: There is no reason to believe that the swims are not independent or that the two groups are not independent of each other.
- Randomization Condition: The swimmers are not a random sample from any identifiable population, but they may be representative of swimmers who tackle challenges such as this.
- Nearly Normal Condition: the boxplots show no outliers. The histograms are unimodal; the histogram for men is somewhat skewed to the right. (Show your graphs.)
37. a)  $H_0: \mu_R - \mu_N = 0$  vs.  $H_A: \mu_R - \mu_N < 0$ .  $t = -1.36$ ,  $df = 20.00$ ,  $P\text{-value} = 0.0945$ . Because the P-value is large, we fail to reject  $H_0$ . These data show no evidence of a difference in mean number of objects recalled between listening to rap or no music at all.
- b) Didn't conclude any difference.

CHAPTER 25

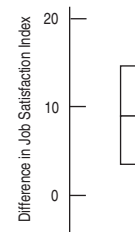
1. a) Randomly assign 50 hens to each of the two kinds of feed. Compare production at the end of the month.
- b) Give all 100 hens the new feed for 2 weeks and the old food for 2 weeks, randomly selecting which feed the hens get first. Analyze the differences in production for all 100 hens.
- c) Matched pairs. Because hens vary in egg production, the matched-pairs design will control for that.
3. a) Show the same people ads with and without sexual images, and record how many products they remember in each group. Randomly decide which ads a person sees first. Examine the differences for each person.
- b) Randomly divide volunteers into two groups. Show one group ads with sexual images and the other group ads without. Compare how many products each group remembers.
5. a) Matched pairs—same cities in different periods.
- b) There is a significant difference ( $P\text{-value} = 0.0244$ ) in the labor force participation rate for women in these cities; women's participation seems to have increased between 1968 and 1972.
7. a) Use the paired  $t$ -test because we have pairs of Fridays in 5 different months. Data from adjacent Fridays within a month may be more similar than data from randomly chosen Fridays.
- b) We conclude that there is evidence ( $P\text{-value} 0.0212$ ) that the mean number of cars found on the M25 motorway on Friday the 13th is less than on the previous Friday.
- c) We don't know if these Friday pairs were selected at random. If these are the Fridays with the largest differences, this will affect our conclusion. The Nearly Normal Condition appears to be met by the differences, but the sample size is small.
9. Adding variances requires that the variables be independent. These price quotes are for the same cars, so they are paired. Drivers quoted high insurance premiums by the local company will be likely to get a high rate from the online company, too.
11. a) The histogram—we care about differences in price.
- b) Insurance cost is based on risk, so drivers are likely to see similar quotes from each company, making the differences relatively smaller.
- c) The price quotes are paired; they were for a random sample of fewer than 10% of the agent's customers; the histogram of differences looks approximately Normal.
13.  $H_0: \mu(\text{Local} - \text{Online}) = 0$  vs.  $H_A: \mu(\text{Local} - \text{Online}) > 0$ ; with 9 df,  $t = 0.83$ . With a high P-value of 0.215, we don't reject the null hypothesis. These data don't provide evidence that online premiums are lower, on average.

15.

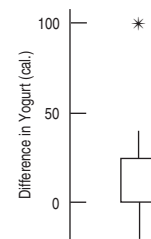


Data are paired for each city; cities are independent of each other; boxplot shows the temperature differences are reasonably symmetric, with no outliers. This is probably not a random sample, so we might be wary of inferring that this difference applies to all European cities. Based on these data, we are 90% confident that the average temperature in European cities in July is between  $32.3^\circ\text{F}$  and  $41.3^\circ\text{F}$  higher than in January.

17. Based on these data, we are 90% confident that boys, on average, can do between 1.6 and 13.0 more push-ups than girls (independent samples—not paired).
19. a) Paired sample test. Data are before/after for the same workers; workers randomly selected; assume fewer than 10% of all this company's workers; boxplot of differences shows them to be symmetric, with no outliers.



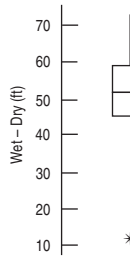
- b)  $H_0: \mu_D = 0$  vs.  $H_A: \mu_D > 0$ .  $t = 3.60$ ,  $P\text{-value} = 0.0029$ . Because  $P < 0.01$ , reject  $H_0$ . These data show evidence that average job satisfaction has increased after implementation of the program.
- c) Type I
21.  $H_0: \mu_D = 0$  vs.  $H_A: \mu_D \neq 0$ . Data are paired by brand; brands are independent of each other; fewer than 10% of all yogurts (questionable); boxplot of differences shows an outlier (100) for Great Value:



With the outlier included, the mean difference (Strawberry – Vanilla) is 12.5 calories with a  $t$ -stat of 1.332, with 11 df, for a P-value of 0.2098. Deleting the outlier, the difference is even smaller, 4.55 calories with a  $t$ -stat of only 0.833 and a P-value of 0.4241. With P-values so large, we do not reject  $H_0$ . We conclude that the data do not provide evidence of a difference in mean calories.

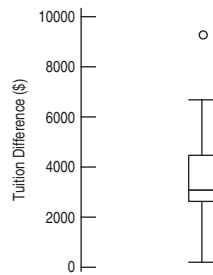
23. a) Cars were probably not a simple random sample, but may be representative in terms of stopping distance; boxplot does not show outliers, but does indicate right skewness. A 95% confidence interval for the mean stopping distance on dry pavement is (131.8, 145.6) feet.

- b) Data are paired by car; cars were probably not randomly chosen, but representative; boxplot shows an outlier (car 4) with a difference of 12. With deletion of that car, a Normal probability plot of the differences is relatively straight.



Retaining the outlier, we estimate with 95% confidence that the average braking distance is between 38.8 and 62.6 feet more on wet pavement than on dry, based on this sample. (Without the outlier, the confidence interval is 47.2 to 62.8 feet.)

25. a) Paired Data Assumption: Data are paired by college. Randomization Condition: This was a random sample of public colleges and universities. 10% Condition: these are fewer than 10% of all public colleges and universities.



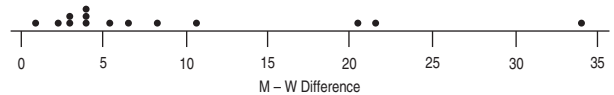
Normal Population Assumption: U.C. Irvine seems to be an outlier; we might consider removing it.

- b) Having deleted the observation for U.C.-Irvine, whose difference of \$9300 was an outlier, we are 90% confident, based on the remaining data, that nonresidents pay, on average, between \$2615.31 and \$3918.02 more than residents. If we retain the outlier, the interval is (\$2759, \$4409).
- c) Assertion is reasonable; with or without the outlier, \$3500 is in the confidence interval.
27. a) 60% is 30 strikes;  $H_0: \mu = 30$  vs.  $H_A: \mu > 30$ .  $t = 6.07$ ,  $P\text{-value} = 3.92 \times 10^{-6}$ . With a very small P-value, we reject  $H_0$ . There is very strong evidence that players can throw more than 60% strikes after training, based on this sample.
- b)  $H_0: \mu_D = 0$  vs.  $H_A: \mu_D > 0$ .  $t = 0.135$ ,  $P\text{-value} = 0.4472$ . With such a high P-value, we do not reject  $H_0$ . These data provide no evidence that the program has improved pitching in these Little League players.

**PART VI REVIEW**

1. a)  $H_0: \mu_{Jan} - \mu_{Jul} = 0$ ;  $H_A: \mu_{Jan} - \mu_{Jul} \neq 0$ .  $t = -1.94$ ,  $df = 43.68$ ,  $P\text{-value} = 0.0590$ . Since  $P < 0.10$ , reject the null. These data show a significant difference in mean age to crawl between January and July babies.
- b)  $H_0: \mu_{Apr} - \mu_{Oct} = 0$ ;  $H_A: \mu_{Apr} - \mu_{Oct} \neq 0$ .  $t = -0.92$ ;  $df = 59.40$ ;  $P\text{-value} = 0.3610$ . Since  $P > 0.10$ , do not reject the null; these data do not show a significant difference between April and October with regard to the mean age at which crawling begins.
- c) These results are not consistent with the claim.

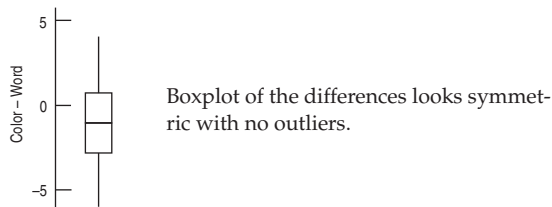
3.  $H_0: p = 0.26$ ;  $H_A: p \neq 0.26$ .  $z = 0.946$ ;  $P\text{-value} = 0.3443$ . Because the P-value is high, we do not reject  $H_0$ . These data do not show that the Denver-area rate is different from the national rate in the proportion of businesses with women owners.
5. Based on these data, we are 95% confident that the mean difference in aluminum oxide content is between  $-3.37$  and  $1.65$ . Since the interval contains 0, the means in aluminum oxide content of the pottery made at the two sites could reasonably be the same.
7. a)  $H_0: p_{ALS} - p_{Other} = 0$ ;  $H_A: p_{ALS} - p_{Other} > 0$ .  $z = 2.52$ ;  $P\text{-value} = 0.0058$ . With such a low P-value, we reject  $H_0$ . This is strong evidence that there is a higher proportion of varsity athletes among ALS patients than those with other disorders.
- b) Observational retrospective study. To make the inference, one must assume the patients studied are representative.
9.  $H_0: \mu = 7.41$ ;  $H_A: \mu \neq 7.41$ .  $t = 2.18$ ;  $df = 111$ ;  $P\text{-value} = 0.0313$ . With such a low P-value, we reject  $H_0$ . Assuming that Missouri babies fairly represent the United States, these data suggest that American babies are different from Australian babies in birth weight; it appears American babies are heavier, on average.
11. a) If there is no difference in the average fish sizes, the chance of seeing an observed difference this large just by natural sampling variation is less than 0.1%.
- b) If cost justified, feed them a natural diet. c) Type I
13. a) Assuming the conditions are met, from these data we are 95% confident that patients with cardiac disease average between 3.39 and 5.01 years older than those without cardiac disease.
- b) Older patients are at greater risk from a variety of other health issues, and perhaps more depressed.
15. a) Stratified sample survey.
- b) We are 95% confident that the proportion of boys who play computer games is between 7.0 and 17.0 percentage points higher than among girls.
- c) Yes. The entire interval lies above 0.
17. Based on the data, we are 95% confident that the mean difference in words misunderstood is between  $-3.76$  and  $3.10$ . Because 0 is in the confidence interval, we would conclude that the two tapes could be equivalent.
19. a)



The countries that appear to be outliers are Spain, Italy, and Portugal. They are all Mediterranean countries.

- b)  $H_0: \mu_D = 0$ ;  $H_A: \mu_D > 0$ .  $t = 5.56$ ;  $df = 10$ ;  $P\text{-value} = 0.0001$ . With such a low P-value, we reject  $H_0$ . These data show that European men are more likely than women to read newspapers.
21. We are 95% confident that the proportion of American adults who would agree with the statement is between 57.0% and 63.0%.
23. Data are matched pairs (before and after for the same rooms); less than 10% of all rooms in a large hotel; uncertain how these rooms were selected (are they representative?). Histogram shows that differences are roughly unimodal and symmetric, with no outliers. A 95% confidence interval for the difference, before - after, is (0.58, 2.65) counts. Since the entire interval is above 0, these data suggest that the new air-conditioning system was effective in reducing average bacteria counts.
25. a) We are 95% confident that between 19.77% and 38.66% of children with bipolar symptoms will be helped with medication and psychotherapy, based on this study.
- b) 221 children
27. a) From this histogram, about 115 loaves or more. (Not Normal.) This assumes the last 100 days are typical.
- b) Large sample size; CLT says  $\bar{y}$  will be approximately Normal.

- c) From the data, we are 95% confident that on average the bakery will sell between 101.2 and 104.8 loaves of bread a day.  
 d) 25  
 e) Yes, 100 loaves per day is too low—the entire confidence interval is above that.
29. a)  $H_0: p_{\text{High}} - p_{\text{Low}} = 0$ ;  $H_A: p_{\text{High}} - p_{\text{Low}} \neq 0$ .  $z = -3.57$ ; P-value = 0.0004. Because the P-value is so low, we reject  $H_0$ . These data suggest the IRS risk is different in the two groups; it appears people who consume dairy products often have a lower risk, on average.  
 b) Doesn't indicate causality; this is not an experiment.
31. Based on these data, we are 95% confident that seeded clouds will produce an average of between  $-4.76$  and  $559.56$  more acre-feet of rain than unseeded clouds. Since the interval contains negative values, it may be that seeding is unproductive.
33. a) Randomizing order of the tasks helps avoid bias and memory effects. Randomizing the cards helps avoid bias as well.  
 b)  $H_0: \mu_D = 0$ ;  $H_A: \mu_D \neq 0$



$t = -1.70$ ; P-value = 0.0999; do not reject  $H_0$ , because  $P > 0.05$ . The data do not provide evidence that the color or written word dominates.

35. a) Different samples give different means; this is a fairly small sample. The difference may be due to natural sampling variation.  
 b)  $H_0: \mu = 100$ ;  $H_A: \mu < 100$   
 c) Batteries selected are a SRS (representative); fewer than 10% of the company's batteries; lifetimes are approximately Normal.  
 d)  $t = -1.0$ ; P-value = 0.1666; do not reject  $H_0$ . This sample does not show that the average life of the batteries is significantly less than 100 hours.  
 e) Type II.

## CHAPTER 26

1. a) Chi-square test of independence. We have one sample and two variables. We want to see if the variable *Account Type* is independent of the variable *Trade Type*.  
 b) Other test. *Account Size* is quantitative, not counts.  
 c) Chi-square test of homogeneity. We want to see if the distribution of one variable, *Courses*, is the same for two groups (resident and nonresident students).
3. a) 10 b) Goodness-of-fit  
 c)  $H_0$ : The die is fair (all faces have  $p = 1/6$ ).  
 $H_A$ : The die is not fair.  
 d) Count data; rolls are random and independent; expected frequencies are all bigger than 5.  
 e) 5 f)  $\chi^2 = 5.600$ , P-value = 0.3471  
 g) Because the P-value is high, do not reject  $H_0$ . The data show no evidence that the die is unfair.
5. a) Weights are quantitative, not counts.  
 b) Count the number of each kind of nut, assuming the company's percentages are based on counts rather than weights.
7.  $H_0$ : The police force represents the population (29.2% white, 28.2% black, etc.).  $H_A$ : The police force is not representative of the population.  $\chi^2 = 16516.88$ ,  $df = 4$ , P-value = 0.0000. Because the P-value is so low, we reject  $H_0$ . These data show that the police force is not representative of the population. In particular, there are too many white officers in relationship to their membership in the community.

9. a)  $\chi^2 = 5.671$ ,  $df = 3$ , P-value = 0.1288. With a P-value this high, we fail to reject  $H_0$ . Yes, these data are consistent with those predicted by genetic theory.  
 b)  $\chi^2 = 11.342$ ,  $df = 3$ , P-value = 0.0100. Because of the low P-value, we reject  $H_0$ . These data provide evidence that the distribution is not as specified by genetic theory.  
 c) With small samples, many more data sets will be consistent with the null hypothesis. With larger samples, small discrepancies will show evidence against the null hypothesis.
11. a)  $96/16 = 6$  b) Goodness of Fit  
 c)  $H_0$ : The number of large hurricanes remains constant over decades.  
 $H_A$ : The number of large hurricanes has changed.  
 d) 15 e) P-value = 0.63  
 f) The very high P-value means these data offer no evidence that the numbers of large hurricanes has changed.  
 g) The final period is only 6 years rather than 10 and already 7 large hurricanes have been observed. Perhaps this decade will have an unusually large number of such hurricanes.
13. a) Independence  
 b)  $H_0$ : Breastfeeding success is independent of having an epidural.  
 $H_A$ : There's an association between breastfeeding success and having an epidural.
15. a) 1 b) 159.34  
 c) Breastfeeding behavior should be independent for these babies. They are fewer than 10% of all babies; we assume they are representative. We have counts, and all the expected counts are at least 5.
17. a) 5.90 b) P-value < 0.005  
 c) The P-value is very low, so reject the null. There's evidence of an association between having an epidural and subsequent success in breastfeeding.  
 $(190 - 159.34) / \sqrt{159.34} = 2.43$
19. a)  $\frac{190 - 159.34}{\sqrt{159.34}} = 2.43$   
 b) It appears that babies whose mothers had epidurals during childbirth are much less likely to be breastfeeding 6 months later.
21. These factors would not be mutually exclusive. There would be yes or no responses for every baby for each.
23. a) 40.2% b) 8.1% c) 62.2% d) 285.48  
 e)  $H_0$ : Survival was independent of status on the ship.  
 $H_A$ : Survival depended on the status.  
 f) 3  
 g) We reject the null hypothesis. Survival depended on status. We can see that first-class passengers were more likely to survive than passengers of any other class.
25. First class passengers were most likely to survive, while 3<sup>rd</sup>-class passengers and crew were under-represented among the survivors.
27. a) Experiment—actively imposed treatments (different drinks)  
 b) Homogeneity  
 c)  $H_0$ : The rate of urinary tract infection is the same for all three groups.  $H_A$ : The rate of urinary tract infection is different among the groups.  
 d) Count data; random assignment to treatments; all expected frequencies larger than 5.  
 e) 2 f)  $\chi^2 = 7.776$ , P-value = 0.020.  
 g) With a P-value this low, we reject  $H_0$ . These data provide reasonably strong evidence that there is a difference in urinary tract infection rates between cranberry juice drinkers, lactobacillus drinkers, and the control group.  
 h) The standardized residuals are

|              | Cranberry | Lactobacillus | Control  |
|--------------|-----------|---------------|----------|
| Infection    | -1.87276  | 1.19176       | 0.68100  |
| No Infection | 1.24550   | -0.79259      | -0.45291 |

From the standardized residuals (and the sign of the residuals), it appears those who drank cranberry juice were less likely to develop urinary tract infections; those who drank lactobacillus were more likely to have infections.

29. a) Independence  
 b)  $H_0$ : *Political Affiliation* is independent of *Sex*.  
 $H_A$ : There is a relationship between *Political Affiliation* and *Sex*.  
 c) Counted data; probably a random sample, but can't extend results to other states; all expected frequencies greater than 5.  
 d)  $\chi^2 = 4.851$ ,  $df = 2$ , P-value = 0.0884.  
 e) Because of the high P-value, we do not reject  $H_0$ . These data do not provide evidence of a relationship between *Political Affiliation* and *Sex*.
31.  $H_0$ : *Political Affiliation* is independent of *Region*.  $H_A$ : There is a relationship between *Political Affiliation* and *Region*.  $\chi^2 = 13.849$ ,  $df = 4$ , P-value = 0.0078. With a P-value this low, we reject  $H_0$ . *Political Affiliation* and *Region* are related. Examination of the residuals shows that those in the West are more likely to be Democrat than Republican; those in the Northeast are more likely to be Republican than Democrat.
33. a) Homogeneity  
 b)  $H_0$ : The grade distribution is the same for both professors.  
 $H_A$ : The grade distributions are different.

|   | Dr. Alpha | Dr. Beta |
|---|-----------|----------|
| A | 6.667     | 5.333    |
| B | 12.778    | 10.222   |
| C | 12.222    | 9.778    |
| D | 6.111     | 4.889    |
| F | 2.222     | 1.778    |

Three cells have expected frequencies less than 5.

35. a) 

|         | Dr. Alpha | Dr. Beta |
|---------|-----------|----------|
| A       | 6.667     | 5.333    |
| B       | 12.778    | 10.222   |
| C       | 12.222    | 9.778    |
| Below C | 8.333     | 6.667    |

  
 All expected frequencies are now larger than 5.  
 b) Decreased from 4 to 3.  
 c)  $\chi^2 = 9.306$ , P-value = 0.0255. Because the P-value is so low, we reject  $H_0$ . The grade distributions for the two professors are different. Dr. Alpha gives fewer A's and more grades below C than Dr. Beta.
37.  $\chi^2 = 14.058$ ,  $df = 1$ , P-value = 0.0002. With a P-value this low, we reject  $H_0$ . There is evidence of racial steering. Blacks are much less likely to rent in Section A than Section B.
39. a)  $z = 3.74936$ ,  $z^2 = 14.058$ .  
 b) P-value ( $z$ ) = 0.0002 (same as in Exercise 25).
41.  $\chi^2 = 5.89$ ,  $df = 3$ , P = 0.117. Because the P-value is  $>0.05$ , these data show no evidence of an association between the mother's age group and the outcome of the pregnancy.

**CHAPTER 27**

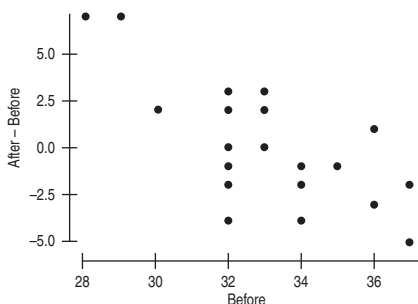
1. a)  $\widehat{Error} = 453.22 - 8.37 YearSince1970$ ; according to the model, the error made in predicting a hurricane's path was about 453 nautical miles, on average, in 1970. It has been declining at a rate of about 8.37 nautical miles per year.  
 b)  $H_0$ :  $\beta_1 = 0$ ; there has been no change in prediction accuracy.  
 $H_A$ :  $\beta_1 \neq 0$ ; there has been a change in prediction accuracy.  
 c) With a P-value  $< 0.001$ , I reject the null hypothesis and conclude that prediction accuracies have in fact been changing during this period.  
 d) 58.5% of the variation in hurricane prediction accuracy is accounted for by this linear model on time.

3. a)  $\widehat{Budget} = -31.387 + 0.714 RunTime$ . The model suggests that movies cost about \$714,000 per minute to make.  
 b) A negative starting value makes no sense, but the P-value of 0.07 indicates that we can't discern a difference between our estimated value and zero. The statement that a movie of zero length should cost \$0 makes sense.  
 c) Amounts by which movie costs differ from predictions made by this model vary, with a standard deviation of about \$33 million.  
 d) 0.154 \$m/min  
 e) If we constructed other models based on different samples of movies, we'd expect the slopes of the regression lines to vary, with a standard deviation of about \$154,000 per minute.
5. a) The scatterplot looks straight enough, the residuals look random and nearly normal, and the residuals don't display any clear change in variability.  
 b) I'm 95% confident that the cost of making longer movies increases at a rate of between 0.41 and 1.02 million dollars per additional minute.
7. a)  $H_0$ :  $\beta_1 = 0$ ; there's no association between calories and sodium content in all-beef hot dogs.  $H_A$ :  $\beta_1 \neq 0$ : there is an association.  
 b) Based on the low P-value (0.0018), I reject the null. There is evidence of an association between the number of calories in all-beef hot dogs and their sodium contents.
9. a) Among all-beef hot dogs with the same number of calories, the sodium content varies, with a standard deviation of about 60 mg.  
 b) 0.561 mg/cal  
 c) If we tested many other samples of all-beef hot dogs, the slopes of the resulting regression lines would be expected to vary, with a standard deviation of about 0.56 mg of sodium per calorie.
11. I'm 95% confident that for every additional calorie, all-beef hot dogs have, on average, between 1.07 and 3.53 mg more sodium.
13. a)  $H_0$ : Difference in age at first marriage has not been changing,  $\beta_1 = 0$ .  $H_A$ : Difference in age at first marriage has been changing,  $\beta_1 \neq 0$ .  
 b) Residual plot shows no obvious pattern; histogram is not particularly Normal, but shows no obvious skewness or outliers.  
 c)  $t = -7.04$ , P-value  $< 0.0001$ . With such a low P-value, we reject  $H_0$ . These data show evidence that difference in age at first marriage is decreasing.
15. Based on these data, we are 95% confident that the average difference in age at first marriage is decreasing at a rate between 0.039 and 0.021 years per year.
17. a)  $H_0$ : *Fuel Economy* and *Weight* are not (linearly) related,  $\beta_1 = 0$ .  $H_A$ : *Fuel Economy* changes with *Weight*,  $\beta_1 \neq 0$ . P-value  $< 0.0001$ , indicating strong evidence of an association.  
 b) Yes, the conditions seem satisfied. Histogram of residuals is unimodal and symmetric; residual plot looks OK, but some "thickening" of the plot with increasing values.  
 c)  $t = -12.2$ , P-value  $< 0.0001$ . These data show evidence that *Fuel Economy* decreases with the *Weight* of the car.
19. a)  $(-9.57, -6.86)$  mpg per 1000 pounds.  
 b) Based on these data, we are 95% confident that *Fuel Efficiency* decreases between 6.86 and 9.57 miles per gallon, on average, for each additional 1000 pounds of *Weight*.
21. a) We are 95% confident that 2500-pound cars will average between 27.34 and 29.07 miles per gallon.  
 b) Based on the regression, a 3450-pound car will get between 15.44 and 25.36 miles per gallon, with 95% confidence.
23. a) Yes.  $t = 2.73$ , P-value = 0.0079. With a P-value so low, we reject  $H_0$ . There is a positive relationship between *Calories* and *Sodium* content.  
 b) No.  $R^2 = 9\%$  and  $s$  appears to be large, although without seeing the data, it is a bit hard to tell.
25. Plot of *Calories* against *Fiber* does not look linear; the residuals plot also shows increasing variance as predicted values get large. The histogram of residuals is right skewed.

27. a)  $H_0$ : No (linear) relationship between *BCI* and *pH*,  $\beta_1 = 0$ .  
 $H_A$ : There seems to be a relationship,  $\beta_1 \neq 0$ .  
 b)  $t = -7.73$  with 161 df; P-value  $< 0.0001$   
 c) There seems to be a negative relationship; *BCI* decreases as *pH* increases at an average of 197.7 *BCI* units per increase of 1 *pH*.
29. a)  $H_0$ : No linear relationship between *Population* and *Ozone*,  $\beta_1 = 0$ .  $H_A$ : *Ozone* increases with *Population*,  $\beta_1 > 0$ .  $t = 3.48$ , P-value = 0.0018. With a P-value so low, we reject  $H_0$ . These data show evidence that *Ozone* increases with *Population*.  
 b) Yes, *Population* accounts for 84% of the variability in *Ozone* level, and *s* is just over 5 parts per million.
31. a) Based on this regression, each additional million residents corresponds to an increase in average ozone level of between 3.29 and 10.01 ppm, with 90% confidence.  
 b) The mean *Ozone* level for cities with 600,000 people is between 18.47 and 27.29 ppm, with 90% confidence.
33. a) 33 batteries.  
 b) Yes. The scatterplot is roughly linear with lots of scatter; plot of residuals vs. predicted values shows no overt patterns; Normal probability plot of residuals is reasonably straight.  
 c)  $H_0$ : No linear relationship between *Cost* and *Cranking Amps*,  $\beta_1 = 0$ .  $H_A$ : *Cranking Amps* increase with cost,  $\beta_1 > 0$ .  $t = 3.23$ ; P-value =  $\frac{1}{2}(0.0029) = 0.00145$ . With a P-value so low, we reject  $H_0$ . These data provide evidence that more expensive batteries do have more cranking amps.  
 d) No.  $R^2 = 25.2\%$  and  $s = 116$  amps. Since the range of amperage is only about 400 amps, an *s* of 116 is not very useful.  
 e)  $\widehat{\text{Cranking amps}} = 384.59 + 4.15 \times \text{Cost}$ .  
 f) (1.97, 6.32) cold cranking amps per dollar.  
 g) *Cranking amps* increase, on average, between 1.97 and 6.32 per dollar of battery *Cost* increase, with 90% confidence.
35. a)  $H_0$ : No linear relationship between *Waist size* and *%Body Fat*,  $\beta_1 = 0$ .  $H_A$ : *%Body Fat* changes with *Waist size*,  $\beta_1 \neq 0$ .  $t = 8.14$ ; P-value  $< 0.0001$ . There's evidence that *%Body Fat* seems to increase with *Waist size*.  
 b) With 95% confidence, mean *%Body Fat* for people with 40-inch waists is between 23.58 and 29.02, based on this regression.
37. a) The regression model is  $\widehat{\text{Midterm2}} = 12.005 + 0.721 \text{ Midterm1}$

|                | Estimate | Std Error | t-ratio  | P-value  |
|----------------|----------|-----------|----------|----------|
| Intercept      | 12.00543 | 15.9553   | 0.752442 | 0.454633 |
| Slope          | 0.72099  | 0.183716  | 3.924477 | 0.000221 |
| <b>RSquare</b> | 0.198982 |           |          |          |
| <b>s</b>       | 16.78107 |           |          |          |
| <b>n</b>       | 64       |           |          |          |

- b) The scatterplot shows a weak, somewhat linear, positive relationship. There are several outlying points, but removing them only makes the relationship slightly stronger. There is no obvious pattern in the residual plot. The regression model appears appropriate. The small P-value for the slope shows that the slope is statistically distinguishable from 0 even though the  $R^2$  value of 0.199 suggests that the overall relationship is weak.  
 c) No. The  $R^2$  value is only 0.199 and the value of *s* of 16.8 points indicates that she would not be able to predict performance on *Midterm2* very accurately.
39.  $H_0$ : Slope of *Effectiveness vs Initial Ability* = 0;  $H_A$ : Slope  $\neq 0$



- Scatterplot is straight enough. Regression conditions appear to be met.  $t = -4.34$ ,  $df = 19$ , P-value = 0.004. With a P-value this small, we reject the null hypothesis. There is strong evidence that the effectiveness of the video depends on the player's initial ability. The negative slope observed that the method is more effective for those whose initial performance was poorest and less so for those whose initial performance was better. This looks like a case of regression to the mean. Those who were above average initially tended to be worse after training. Those who were below average initially tended to improve.
41. a) Data plot looks linear; no overt pattern in residuals; histogram of residuals roughly symmetric and unimodal.  
 b)  $H_0$ : No linear relationship between *Education* and *Mortality*,  $\beta_1 = 0$ .  $H_A$ :  $\beta_1 \neq 0$ .  $t = -6.24$ ; P-value  $< 0.001$ . There is evidence that cities in which the mean education level is higher also tend to have a lower mortality rate.  
 c) No. Data are on cities, not individuals. Also, these are observational data. We cannot predict causal consequences from them.  
 d) (-65.95, -33.89) deaths per 100,000 people.  
 e) *Mortality* decreases, on average, between 33.89 and 65.95 deaths per 100,000 for each extra year of average *Education*.  
 f) Based on the regression, the average *Mortality* for cities with an average of 12 years of *Education* will be between 874.239 and 914.196 deaths per 100,000 people.

## PART VII REVIEW

- $H_0$ : The proportions are as specified by the ratio 1:3:3:9;  $H_A$ : The proportions are not as stated.  $\chi^2 = 5.01$ ;  $df = 3$ ; P-value = 0.1711. Since  $P > 0.05$ , we fail to reject  $H_0$ . These data do not provide evidence to indicate that the proportions are other than 1:3:3:9.
- a)  $H_0$ : *Mortality* and *calcium concentration* in water are not linearly related,  $\beta_1 = 0$ ;  $H_A$ : They are linearly related,  $\beta_1 \neq 0$ .  
 b)  $t = -6.73$ ; P-value  $< 0.0001$ . There is a significant negative relationship between calcium in drinking water and mortality.  
 c) (-4.19, -2.27) deaths per 100,000 for each ppm calcium.  
 d) Based on the regression, we are 95% confident that mortality (deaths per 100,000) decreases, on average, between 2.27 and 4.19 for each part per million of calcium in drinking water.
- 404 checks
- $H_0$ : *Income* and *Party* are independent.  $H_A$ : *Income* and *Party* are not independent.  $\chi^2 = 17.19$ ; P-value = 0.0018. With such a small P-value, we reject  $H_0$ . These data show evidence that income level and party are not independent. Examination of components suggests Democrats are most likely to have low incomes; Independents are most likely to have middle incomes, and Republicans are most likely to have high incomes.
- $H_0$ :  $p_L - p_R = 0$ ;  $H_A$ :  $p_L - p_R \neq 0$ .  $z = 1.38$ ; P-value = 0.1683. Since  $P > 0.05$ , we do not reject  $H_0$ . These data do not provide evidence of a difference in musical abilities between right- and left-handed people.
- a)  $H_0$ :  $\mu_D = 0$ ;  $H_A$ :  $\mu_D \neq 0$ .  
 Boxplot of the differences indicates a strong outlier (1958). With the outlier kept in, the *t*-stat is 0, with a P-value of 1.00 (two sided). There is no evidence of a difference (on average of actual and that predicted by Gallup. With the outlier taken out, the *t*-stat is still only -0.8525 with a P-value of 0.4106, so the conclusion is the same.  
 b)  $H_0$ : There is no (linear) relationship between predicted and actual number of Democratic seats won ( $\beta_1 = 0$ ).  $H_A$ : There is a relationship ( $\beta_1 \neq 0$ ). The relationship is very strong, with an  $R^2$  of 97.7%. The *t*-stat is 22.56. Even with only 12 df, this is clearly significant (P-value  $< 0.0001$ ). There is an outlying residual (1958), but without it, the regression is even stronger.
- Conditions are met;  $df = 4$ ;  $\chi^2 = 0.69$ ; P-value = 0.9526. Since  $P > 0.05$ , we do not reject  $H_0$ . We do not have evidence that the way the hospital deals with twin pregnancies has changed.



15. a) Based on these data, the average annual rainfall in LA is between 11.65 and 17.39 inches, with 90% confidence.  
 b) About 46 years  
 c) No. The regression equation is  $\widehat{Rain} = -51.684 + 0.033 \times Year$ .  $R^2 = 0.1\%$ . For the slope,  $t = 0.12$  with P-value = 0.9029.
17. a) Linear regression is meaningless—the data are categorical.  
 b) This is a two-way table that is appropriate.  $H_0$ : *Eye* and *Hair* color are independent.  $H_A$ : *Eye* and *Hair* color are not independent. However, four cells have expected counts less than 5, so the  $\chi^2$  analysis is not valid unless cells are merged. However, with a  $\chi^2$  value of 223.6 with 16 df and a P-value < 0.0001, the results are not likely to change if we merge appropriate eye colors.
19. a)  $H_0: p_Y - p_O = 0$ ;  $H_A: p_Y - p_O \neq 0$ .  $z = 3.56$ ; P-value = 0.0004. With such a small P-value, we reject  $H_0$ . We conclude there is evidence of a difference in effectiveness; it appears the methods are not as good for older women.  
 b)  $\chi^2 = 12.70$ ; P-value = 0.0004. Same conclusion.  
 c) The P-values are the same;  $z^2 = (3.563944)^2 = 12.70 = \chi^2$ .
21. a) Positive direction, generally linear trend; moderate scatter.  
 b)  $H_0$ : There is no linear relationship between *Interval* and *Duration*.  $\beta_1 = 0$ .  $H_A$ : There is a linear relationship,  $\beta_1 \neq 0$ .  
 c) Yes; histogram is unimodal and roughly symmetric; residuals plot shows random scatter.  
 d)  $t = 27.1$ ; P-value  $\leq 0.001$ . With such a small P-value, we reject  $H_0$ . There is evidence of a positive linear relationship between duration and time to next eruption of Old Faithful.  
 e) The average time to next eruption after a 2-minute eruption is between 53.24 and 56.12 minutes, with 95% confidence.  
 f) Based on this regression, we will have to wait between 63.23 and 87.57 minutes after a 4-minute eruption, with 95% confidence.
23. a)  $t = 1.42$ , df = 459.3, P-value = 0.1574. Since  $P > 0.05$ , we do not reject  $H_0$ . There's no evidence the two groups differed in ability at the start of the study.  
 b)  $t = 15.11$ ; P-value < 0.0001. The group taught using the accelerated Math program showed a significant improvement.  
 c)  $t = 9.24$ ; P-value < 0.0001. The control group showed a significant improvement in test scores.  
 d)  $t = 5.78$ ; P-value < 0.0001. The Accelerated Math group had significantly higher gains than the control group.
25. a) The regression—he wanted to know about association.  
 b) There is a moderate relationship between cottage cheese and ice cream sales; for every million pounds of cottage cheese, 1.19 million pounds of ice cream are sold, on average.  
 c) Testing if the mean difference is 0 (matched  $t$ -test). Regression won't answer this question.  
 d) The company sells more cottage cheese than ice cream, on average.  
 e) part (a)—linear relationship; residuals have a Normal distribution; residuals are independent with equal variation about the line. (c)—Observations are independent; differences are approximately Normal; less than 10% of all possible months' data.  
 f) About 71.32 million pounds. g) (0.09, 2.29)  
 h) From this regression, every million pounds of cottage cheese sold is associated with an increase in ice cream sales of between 0.09 and 2.29 million pounds.
27. Based on these data, the average weight loss for the clinic is between 8.24 and 10.06 pounds, with 95% confidence. The clinic's claim is plausible.
29.  $\chi^2 = 8.23$ ; P-value = 0.0414. There is evidence of an association between *cracker type* and *bloating*. Standardized residuals for the gum cracker are  $-1.32$  and  $1.58$ . Prospects for marketing this cracker are not good.

# Photo Acknowledgments

**Page 2** ©AP/Wideworld **Page 3** ©United Features Syndicate **Page 7** ©Dorling Kindersley  
**Page 7** ©AP/Wideworld **Page 10** ©Photoyear (RF) **Page 13** ©Eric Gaillard/Reuters/Corbis  
**Page 20** ©PhotoFest **Page 21** ©The Florence Nightingale Museum **Page 24** ©AP/Wideworld  
**Page 31** ©Stone/Getty Images **Page 44** ©Getty Images **Page 51** ©Corbis Sygma **Page 52** ©MGM/  
 PhotoFest **Page 56** ©Digital Vision **Page 80** ©PhotoDisc **Page 87** Courtesy of Beth Anderson  
**Page 104** ©Jeff Haynes/AFP/Getty Images **Page 105** Paul Kane/Getty Images, Inc. **Page 146**  
 Courtesy of National Oceanic and Atmospheric Administration **Page 161** ©AP Wideworld Photos  
**Page 171** ©Landov Photo **Page 174** ©Bettmann/Corbis **Page 178** ©Digital Vision **Page 201**  
 Courtesy of the authors **Page 204** ©Universal Press Syndicate **Page 205 (Ralph Nader)** ©Getty  
 Images **Page 205 (Pat Buchanan)** ©AP/Wideworld Photo **Page 210** ©PhotoDisc **Page 222**  
 ©Getty Images **Page 255** ©PhotoDisc **Page 256** ©United Feature Syndicate, Inc. **Page 257 (Tiger  
 Woods)** ©Getty Images **Page 257 (Serena Williams)** ©Reuters/Corbis **Page 257 (David Beckham)**  
 ©Getty Sports **Page 268** ©Digital Vision **Page 269 (George Gallup)** ©AP Wideworld Photos  
**Page 276** ©FoodPix/Jupiter Images **Page 279** ©Universal Press Syndicate **Page 281** ©Creators  
 Syndicate **Page 292** Courtesy of Beth Anderson **Page 293** ©iStockphoto **Page 294** ©University of  
 St Andrews MacTutor History of Mathematics archive **Page 296** ©University of St Andrews  
 MacTutor History of Mathematics archive **Page 296** ©United Feature Syndicate, Inc. **Page 297**  
**(dogs)** ©Digital Vision **Page 297 (tomatoes)** ©Dorling Kindersley **Page 303** By permission of  
 John L. Hart FLP and Creators Syndicate, Inc. **Page 305** ©iStockphoto **Page 307** ©Shutterstock  
**Page 324** ©Shutterstock **Page 326** ©University of St Andrews MacTutor History of Mathematics  
 archive **Page 327** ©PhotoEdit **Page 328** ©PhotoDisc Red **Page 329** ©University of St Andrews  
 MacTutor History of Mathematics archive **Page 341** ©Getty News & Sport **Page 342 (colored  
 glass)** ©Corbis Royalty Free **Page 342 (dollar bills)** Courtesy of the US Department of Treasury  
**Page 358** ©St. Andrew's University History of Mathematics Archive **Page 366** ©Digital Vision  
**Page 376** ©University of Texas **Page 378** ©PhotoDisc **Page 388 (conveyor belt of jacks)**  
 ©Stephen Mallon/Imagebank/Getty Images **Page 388 (Tiger Woods)** ©Getty Images **Page 389**  
**(Bernoulli)** ©St. Andrew's University History of Mathematics Archive **Page 389 (Calvin and  
 Hobbes)** ©Universal Press Syndicate **Page 397** ©PhotoDisc **Page 408** Courtesy of Richard De  
 Veaux **Page 412** Dejan Lazarevic/Shutterstock **Page 413** ©St. Andrew's University MacTutor  
 Archive **Page 439** ©PhotoDisc **Page 444** ©Universal Press Syndicate **Page 459** ©Adam Woolfitt/  
 Corbis **Page 480** ©Getty Royalty Free **Page 484** ©Courtesy of Beth Anderson **Page 486** ©University  
 of York Department of Mathematics **Page 488** Courtesy of the Massachusetts Governor's Highway  
 Safety Bureau **Page 504** ©Digital Vision **Page 505** ©The Kobal Collection **Page 511** ©PhotoDisc  
**Page 530** ©Digital Vision **Page 533** ©University of York Department of Mathematics **Page 560**  
 ©Corbis Royalty Free **Page 573** ©PhotoDisc, Corbis Royalty Free and Beth Anderson **Page 574**  
 ©PhotoDisc **Page 587** ©Getty Editorial **Page 589** Courtesy of David Bock **Page 618** ©Royalty-Free/  
 Corbis **Page 626** ©PhotoDisc **Page 628** ©PhotoDisc **Page 632** ©PhotoDisc **Page 649** ©Stone/  
 Getty Images **Page 653** ©PhotoDisc **Page 661** ©Nenana Ice Classic **Page 687** Courtesy of David  
 Bock **Page 28–1 (on DVD)** ©PhotoDisc **Page 28–20 (on DVD)** ©St. Andrew's University MacTutor  
 Archive **Page 29–1 (on DVD)** ©PhotoDisc

Note: Page numbers in **boldface** indicate chapter-level topics; page numbers in *italics* indicate definitions; FE indicates For Example references.

- 5-number summary, 56  
 boxplots and, 81–82, 109  
 10% Condition  
   Central Limit Theorem, 422  
   for chi-square tests, 628  
   for comparing means, 563  
   comparing proportions, 506  
   for confidence intervals, 446  
   independence and, 391  
   for paired data, 590  
   sampling distribution models, 415  
   for Student's *t*-models, 536–537  
 68-95-99.7 Rule, 113  
   Central Limit Theorem and, 414  
   symmetric distribution and, 224  
   working with, 114FE–115FE
- A**
- ActivStats Multimedia Assistant, 5  
 Actuaries, 366  
 Addition Rule, 330–331  
   applying, 331FE, 393  
   General Addition Rule, 342–343, 345FE–346FE  
   for variances, 373  
 Adjusted  $R^2$ , 29-16  
 Agresti-Coull interval, 490FE–491FE  
 Alpha levels, 486, 486–487, 496–497, 547  
 Alternative hypothesis, 460, 463  
   many-sided alternative, 626  
   one-sided alternative, 466, 481FE, 485  
   two-sided alternative, 466  
 Amazon.com, 7–8  
 American Association for Public Opinion Research (AAPOR), 350  
*American Journal of Health Behavior*, 354  
 Analysis of Variance, *See* ANOVA (Analysis of Variance)  
 Annenberg Foundation, 427  
 ANOVA (Analysis of Variance), **28-1–28-40**, 28-6  
   assumptions and conditions, 28-13–28-15  
   balance, 28-18  
   Bonferroni multiple comparisons, 28-19–28-21  
   boxplots for, 28-13  
   common problems, 28-24–28-25  
   comparing means, 28-18–28-19  
   comparing means of groups, 28-2–28-3  
   on the computer, 28-27  
   Does the Plot Thicken? Condition, 28-14  
   Equal Variance Assumption, 28-14, 28-20  
   Error Mean Square, 28-5  
   *F*-statistic, 28-5–28-6  
   *F*-tables, 28-7–28-8  
   handwashing methods example, 28-1–28-2  
   hot beverage containers example, 28-15FE–28-18FE  
   Independence Assumption, 28-13  
   Nearly Normal Condition, 28-15  
   Normal Population Assumption, 28-15  
   on observational data, 28-21  
   Randomization Condition, 28-13  
   residual standard deviation, 28-12–28-13  
   Similar Spread Condition, 28-14  
   Treatment Mean Square (MST), 28-6  
   TV watching example, 28-22FE–28-23, 28-24  
 ANOVA model, 28-9–28-12  
 ANOVA tables, 28-7, 29-9–29-10  
 Area codes, 9  
 Area principle, 22, 33, 48  
 Armstrong, Lance, 13, 222  
 Association(s), 147  
   between categorical variables, 29FE  
   correlation properties, 156  
   direction of, 147, 152, 156  
   linear, 147, 152, 156, 160  
   looking at, 154FE–155FE  
   vs. correlation, 160  
 Assumptions, 112  
   for ANOVA, 28-13–28-15  
   checking, 112, 507FE  
   for chi-square tests, 627–628, 634  
   common problems, 452  
   in comparing counts, 619–620, 620FE  
   comparing proportions, 506  
   and conditions, 112, 184, 415–416  
   confidence intervals, 446  
   counts, 620FE  
   Equal Variance Assumption, 181, 184, 574, 652, 28-14, 28-20, 29-6  
   Independence Assumption, 332, 415, 422, 446, 506, 536, 563, 589–590, 619–620, 652, 28-13, 29-5  
   Independent Groups Assumption, 506, 563–564, 576  
   Linearity Assumption, 184, 201, 651, 29-5  
   for means, 563–564, 564FE  
   Normal Population Assumption, 537, 563, 590, 652–653, 28-15  
   Normality Assumption, 112, 380, 537, 549–550, 563, 590, 653, 28-15, 29-6–29-7  
   for paired data, 589–590, 590FE  
   Paired Data Assumption, 589  
   for regression, 184, 651–653, 653FE–654FE  
   Sample Size Assumption, 415, 422, 446, 620  
   sampling distribution models, 415–416  
   Student's *t*-models, 536–537, 537FE–538FE
- B**
- Balance, 28-18  
 Bar charts, 22  
   area principle, 22  
   Categorical Data Condition and, 23–24  
   common problems, 65–66  
   relative frequency, 23  
   segmented, 30, 30–33  
   *Titanic* example, 22, 28  
 Batteries, life of, 560  
 Bayes, Thomas, 358  
 Bayes's Rule, 358, 483n  
 Bernoulli, Daniel, 389  
 Bernoulli, Jacob, 326, 389  
 Bernoulli trials, 388  
   Binomial probability model, 392–394, 395FE–396FE  
   common problems, 399  
   geometric probability model, 389  
   independence and, 390–391  
 Berra, Yogi, 147, 326, 331  
 Between Mean Square, 28-5  
 Bias(es), 269  
   common problems, 284, 452, 550  
   nonresponse, 283–284  
   in samples, 269, 274, 282FE, 283FE  
   voluntary response, 282  
 Bill and Melinda Gates Foundation, 427–428  
 Bimodal distribution, 50, 116, 422, 537, 590  
 Binomial probability model, 393  
   in Bernoulli trials, 392–394, 395FE–396FE  
   calculator tips for, 396  
   common problems, 399  
   on the computer, 401  
   deriving mean and standard deviation, 394  
   Normal models and, 398  
   spam example, 395FE, 398FE  
   Success/Failure Condition, 397  
   universal blood donor example, 395FE–396FE  
 Blinding, 301, 302FE  
 Blocking, 296, 304, 598  
   paired data and, 588, 598  
   pet food example, 305FE  
 Blocking variable, 296  
 Body fat measurement, 649–650, 29-1, 29-7FE–29-9FE  
 Bonferroni method, 28-19–28-21, 28-20  
 Boxplots, 81  
   5-number summary and, 81–82, 109  
   for ANOVA, 28-13  
   calculator tips for, 86  
   common problems, 576, 599  
   comparing groups with, 83–84, 86  
   handwashing methods, 28-1–28-2  
   outliers in, 81  
   plotting, 560–561  
   re-expressing data, 224–225  
   wind speed example, 81–84  
 Bozo the clown as outlier, 161, 207  
 Buchanan, Pat, 205–206  
 Burger King menu items, 171–175  
 Bush, George W., 205, 207

## C

- Calculators, *See* Graphing calculators
- Cancer, smoking and, 157–158
- Card shuffling, 257
- Carnegie Corporation, 427
- Carpal tunnel syndrome, 568
- Cases, 9
- Categorical data, 20–43
  - area principle, 22
  - bar charts, 22–23
  - chi square, 620–640
  - common problems, 33–35
  - conditional distributions, 26–29
  - contingency tables, 24–26, 27, 80
  - Counted Data condition, 619, 627–628
  - displaying on computers, 37
  - frequency tables, 21–22
  - pie charts, 23–24
  - proportions and, 419, 422–424
  - rules of data analysis, 21
  - segmented bar charts, 30–33
- Categorical Data Condition, 24
- Categorical variables, 10
  - bar charts, 22–23
  - chi-square and, 620–640
  - correlation and, 160
  - counting, 11–12
  - distribution of, 22
- Causation
  - chi-square tests and, 639
  - common problems, 29–17
  - correlation and, 157, 160
  - lurking variables and, 208–209
- Ceci, Stephen, 306–307
- Cedar Point amusement park, 88FE
- Cells of tables, 24, 619
- Census, 271, 271–272, 538
- Center for Collaborative Education, 427
- Center for School Change, 427
- Center of distributions, 49, 53
  - describing, 60FE
  - flight cancellation example, 56FE–58FE
  - mean and, 58–60
  - median and, 52–54
  - standardizing z-scores, 110
- Centers for Disease Control, 107FE
- Central Limit Theorem (CLT), 421, 531
  - 10% Condition, 422
  - assumptions and conditions, 422–423
  - Independence Assumption, 422
  - inferences for regression, 668
  - Large Enough Sample Condition, 422–423
  - mean and, 531, 531FE–532FE
  - Normal model and, 428
  - Randomization Condition, 422
  - for sample proportions, 412–414
  - Sample Size Assumption, 422
  - sampling distribution models, 421, 421–422, 429, 532
  - standard deviations and, 531
- Chi-square components, 631
- Chi-square models, 621, 631, 639
- Chi-square statistic, 621
  - calculating, 620–621
  - hypothesis testing for, 621, 622FE–623FE, 624
  - P-values, 638FE
  - process for, 623–624
- Chi-square tests
  - 10% Condition, 628
  - assumptions and conditions, 627–628, 634
  - calculations, 628–629
  - calculator tips for, 638
  - causation and, 639
  - on the computer, 642
  - contingency tables and, 642
  - Counted Data Condition, 627–628
  - Expected Cell Frequency Condition, 628
  - for goodness-of-fit, 618–619, 622FE–623FE, 626
  - of homogeneity, 627, 629FE–630FE, 638
  - for independence, 632–633, 633FE–636FE, 639
  - null hypothesis and, 626–640
  - Randomization Condition, 628
  - residuals for, 623, 631, 636–637
  - writing conclusions for, 638FE
- Cluster sample, 275, 275–276, 275FE
- Coefficient(s)
  - common problems, 29–17
  - multiple regression, 29–3–29–4, 29–10–29–11
  - regression, 273
  - t-ratios for, 29–10–29–11
- Complement Rule, 330, 330FE, 346FE, 371
- Completely randomized design, 298FE, 305–306
- Completely randomized experiments, 305
- Computers
  - ANOVA, 28–27
  - checking Nearly Normal Condition, 542
  - chi-square tests, 642
  - comparing distributions, 94
  - confidence intervals for proportions, 454
  - differences between proportions, 519
  - displaying categorical data, 37
  - displaying quantitative data, 71
  - experiments and, 312
  - hypothesis tests, 476, 498–499
  - inference for means, 552–553
  - linear regression, 192
  - Normal probability plots, 129
  - paired t-analyses, 601–602
  - random variables, 383
  - re-expressing data, 239
  - regression analysis, 672, 29–21
  - regression diagnosis, 213
  - sampling on, 287
  - scatterplots and correlation, 163
  - simulations, 264
  - statistics packages, 16
  - two-sample methods, 579
- Condition(s), 112
  - 10% Condition, 391, 415, 422, 446, 506, 536–537, 563, 590, 628
  - for ANOVA, 28–13–28–15
  - Categorical Data Condition, 24, 31FE
  - checking, 464FE, 507FE
  - for chi-square tests, 627–628, 634
  - common problems, 474
  - in comparing counts, 619–620, 620FE
  - comparing proportions, 506
  - confidence intervals, 446
  - correlation, 152–153
  - Counted Data Condition, 619, 627–628
  - Does the Plot Thicken? Condition, 181, 184, 652, 28–14, 29–6
  - Expected Cell Frequency Condition, 620, 628
  - for fitting models, 203
  - for inference in regression, 651–653, 653FE–654FE, 654
  - Large Enough Sample Condition, 422–423
  - for means, 563–564, 564FE
  - Nearly Normal Condition, 112, 126, 537, 542, 550, 563, 590, 652, 28–15, 29–6–29–7, 29–13
  - Outlier Condition, 153, 178, 184, 652
  - for paired data, 589–590, 590FE
  - Quantitative Data Condition, 49
  - Quantitative Variables Condition, 152, 178, 184
  - Randomization Condition, 415, 422, 446, 506, 536, 563, 589–590, 620, 628, 652, 28–13, 29–5
  - sampling distribution models, 415–416
  - Similar Spread Condition, 574, 28–14
  - Straight Enough Condition, 152, 161, 178, 224, 651, 29–5
  - Student's t-models, 536–537, 537FE–538FE
  - Success/Failure Condition, 397, 415–416, 446, 460, 507
- Conditional distribution, 26
  - and conditional probability, 346–348
  - finding, 27FE
  - pie charts of, 27
  - Titanic example, 26–29
- Conditional probability, 347
  - Bayes's Rule, 358
  - common problems, 359
  - conditional distribution of, 346–348
  - contingency tables and, 346, 351
  - DWI test example, 352FE–353FE
  - examples, 342
  - food survey example, 348FE
  - General Addition Rule, 342–343, 345FE–346FE
  - General Multiplication Rule, 348, 355–356
  - for independent events, 349–351
  - independent vs. disjoint events, 350
  - null hypothesis and, 464
  - P-value as, 483–484
  - relative frequencies of, 347
  - reversing, 356, 357FE–358FE, 359FE
  - room draw, 353–354
  - tree diagrams for, 354–356
- Confidence interval(s), 439–458, 441
  - calculator tips for, 448–449, 510, 541, 567–568, 597
  - census and, 538
  - choosing sample size, 449–450, 450FE
  - common problems, 451–452, 551
  - on the computer, 454
  - creating, 567–568, 597
  - critical values, 445
  - for difference in independent means, 565FE
  - for difference in proportions, 507
  - and effect size, 465
  - hypothesis tests and, 487–488, 488FE–490FE, 547
  - interpreting, 441, 541–542
  - making decisions based on, 488FE
  - margin of error, 442–443
  - for matched pairs, 594–595
  - for a mean, 533–535, 541, 542, 575
  - for mean predicted value, 667
  - for means of independent groups, 561–563
  - paired-t confidence interval, 594–595, 595FE–596FE
  - for predicted values, 667, 667FE
  - for proportions, 439–458, 447FE–448FE
  - for regression slope, 660
  - in sampling distribution models, 440–442
  - for small samples, 490, 490FE–491FE
  - in Student's t-models, 534
  - for two-proportion z-interval, 508, 508FE–510FE
- Confounding, 306
  - in experiments, 306–308
  - lurking variable vs., 307–308
  - pet food example, 307FE
- Constants, changing random variables, 372FE–373FE
- Consumer Reports, 9FE, 11FE
- Context for data, 8
- Contingency tables, 24, 633
  - chi-square tests and, 642
  - conditional probability and, 346, 351
  - examining, 31FE–32FE
  - Titanic example, 24–26, 27
  - Venn diagrams and, 351
- Continuity correction, 399n
- Continuous probabilities, 329

- Continuous random variables, 366, 377, 399
- Control groups, 301
- Controlling sources of variation, 295
- Convenience sample, 282, 282–283, 530
- Cornell University, 92, 569
- Correlation, 152
  - association vs., 160
  - calculator tips for, 155
  - categorizing, 156
  - causation and, 157, 160
  - changing scales, 156FE
  - on computers, 163
  - conditions, 152–153
  - least squares line and, 173–174
  - linear association and, 152, 156
  - notation for, 273
  - Outlier Condition, 153
  - outliers and, 153, 156, 158, 161
  - Quantitative Variables Condition, 152 and regression, 171
  - in scatterplots, 150–153
  - Straight Enough Condition, 152
  - straightening scatterplots, 158–159
- Correlation coefficient, 152
  - direction of association and, 156
  - linear association and, 161
  - outliers and, 161
  - properties, 156
- Correlation tables, 158
- Counted Data Condition, 619, 627–628
- Counts, **618–648**
  - 10% Condition, 628
  - assumptions and conditions, 619–620, 620FE, 627–628
  - calculating, 620–621
  - Categorical Data Condition and, 24
  - categorical variables and, 11–12
  - for chi-square model, 621–640, 629FE–630FE
  - common problems, 639–640
  - comparing observed distributions, 626–627
  - Counted Data Condition, 619, 627–628
  - Expected Cell Frequency Condition, 620, 628
  - finding expected, 619FE
  - frequency tables and, 21–22
  - goodness-of-fit tests, 618–619
  - Independence Assumption, 619–620
  - Randomization Condition, 620, 628
  - Sample Size Assumption, 620
- Critical value(s), 445
  - calculator tips for, 536
  - from *F*-model, 28–8
  - from Normal model, 534
  - from Student's *t*-models, 534, 540
- D**
- Dabilis, Andrew, 87
- Data, 7–19, 8
  - calculator tips for, 14–15
  - categorical. *See* Categorical data characteristics about, 9–11
  - common problems, 14
  - context for, 8
  - counting, 11–12
  - identifiers for, 12
  - plotting, 560–561
  - quantitative. *See* Quantitative data rescaling, 108–109
  - shifting, 107–108
- Data analysis
  - displaying quantitative data, 49
  - of outliers, 87–88
  - rules of, 21, 23–24
- Data table, 8
- De Moivre's Rule, *See* 68-95-99.7 Rule
- Degrees of freedom (df), 533
  - chi-square models and, 621
  - Error Mean Square and, 28–6
  - means and, 549
  - Multiple regression and, 29–2
  - paired-*t* and, 591
  - Regression models and, 657
  - Student's *t*-models and, 533–535
  - Treatment Mean Square and, 28–6
  - Two-sample *t* and, 562, 562n, 563
- Delimiters, 16
- Dependent variables, 149n, 640
- Deviation, 60–61
- Diaconis, Persi, 257
- Dice games, 259FE–260FE
- Direction of association, 147, 152, 156
- Discrete random variables, 366, 370FE–371FE
- Disjoint events, 330
  - Addition Rule, 330
  - common problems, 335
  - DWI test example, 352FE–353FE
  - independent vs., 350
  - Probability Assignment Rule, 331
- Distributions, 22, 44
  - 5-number summary, 56
  - bimodal, 50, 116, 422, 537, 590
  - of categorical variables, 22
  - center of. *See* Center of distributions
  - chi-square. *See* Chapter 26
  - common problems, 92
  - comparing, 81
  - comparing groups, 84FE–85FE
  - comparing groups with boxplots, 83–84
  - comparing groups with histograms, 82
  - comparing observed, 626–627
  - comparing on computers, 94
  - conditional, 26, 26–29, 346–348
  - F*. *See* Chapter 28
  - flight cancellation example, 56FE–58FE
  - marginal, 24, 26FE
  - multimodal, 50, 550
  - Normal. *See* Chapter 6
  - outliers in, 87–88
  - quantitative variables, 44
  - re-expressing data, 89–91
  - shapes of. *See* Shapes of distributions
  - skewed, 50
  - spread of. *See* Spread
  - summarizing, 63FE–64FE
  - symmetric, 50, 58–60, 89–91
  - t*. *See* Chapter 23
  - tails of, 50
  - timeplots of, 88–89
  - uniform, 50
  - unimodal, 50, 535–537
  - of variables, 224
  - wind speed example, 80
- Does the Plot Thicken? Condition, 181, 184, 652, 28-14, 29-6
- Dotplots, 49
- Double-blind, 302
- E**
- Earthquakes, 44–45, 52–56
- Education, Department of, 427
- Educational Testing Service (ETS), 110
- Effect size, 493
  - confidence intervals and, 465
  - errors and, 494–495
  - hypothesis testing and, 492
  - for paired data, 597, 598FE
- Empirical probability, 326
- Empirical Rule, *See* 68-95-99.7 Rule
- Equal Variance Assumption
  - for ANOVA, 28-14, 28-20
  - for linear regression, 181, 184, 652
  - for multiple regression, 29-6
  - for pooled *t*-tests, 574
- Error(s)
  - in data collection, 87
  - effect size and, 494–495
  - in extrapolation, 204
  - in retrospective studies, 293
  - sampling, 414
  - standard. *See* Standard error(s)
  - Type I. *See* Type I error
  - Type II. *See* Type II error
- Error Mean Square (MS<sub>E</sub>), 28-5
- Error Sum of Squares, 28-11–28-12
- Events, 325
  - disjoint, 330–331, 335, 352FE–353FE
  - probability of, 326
- Expected Cell Frequency Condition, 620, 628
- Expected value, 367
  - for chi-square statistic, 623, 628, 631
  - common problems, 380
  - of geometric model, 389
  - of random variables, 366–368, 370FE–371FE, 377
  - restaurant discount example, 368FE
- Experiment(s), **292–316**, 294
  - adding factors, 305–306
  - blinding in, 301–302
  - blocking in, 303–304
  - common problems, 308–309
  - completely randomized two-factor, 305
  - computers and, 312
  - confounding in, 306–308
  - diagrams in, 297
  - differences in treatment groups, 299–300
  - factors in, 294
  - lurking variables, 307–308
  - placebos in, 302–303
  - random assignments in, 294, 296
  - response variables in, 294
  - samples and, 300–301
- Experimental design
  - completely randomized, 298FE, 305–306
  - fertilizer example, 297FE–299FE
  - pet food example, 297FE
  - principles of, 295–297
- Experimental units, 294, 296
- Explanatory variables, 149, 294
- Exposed to smoke (ETS), 91
- Extrapolation, 203–205, 204, 669
- F**
- F*-distribution, 28-6
- F*-statistic, 28-6, 28-7, 29-9
- f*/stops, 158–160, 233–234
- F*-tables, 28-7–28-8
- F*-test, 28-6, 28-15, 29-9–29-10
- Factor(s), 294
  - adding to experiments, 305–306
  - confounding and, 307
  - in experiments, 294
  - level of, 294
- False negative, 357FE, 491. *See also* Type II Error
- False positive, 357FE. *See also* Type I Error
- Far outliers, 81
- Farr, William, 21
- FDA (Food and Drug Administration), 295
- Fechner, Gustav, 294
- Fisher, Ronald Aylmer, 157, 486, 495, 536, 28-4, 28-6, 29-9
- Flight cancellations, 56FE–58FE
- Food and Drug Administration (FDA), 295
- Frequency tables, 21, 21–22
- Friendship affecting price, 569

## G

Gallup, George, 269  
 Galton, Francis, 174  
 Gaps in histograms, 45  
 Gastric freezing, 300  
 General Addition Rule, 342–343, 343FE, 345FE–346FE  
 General Multiplication Rule, 348, 355–356  
 Geometric probability model, 389  
   for Bernoulli trials, 389  
   calculator tips for, 392  
   common problems, 399  
   on the computer, 401  
   spam example, 390FE  
   universal blood donor example, 391FE–392FE  
 Ghosts, belief in, 412  
 Ginkgo biloba, 303  
 Golden Ratio, 152n  
 Goodness-of-fit test (chi-square), 618, 618–619, 622FE–623FE, 624–625, 626  
 Gore, Al, 205–206  
 Gosset, William S., 532–533  
 Grading on a curve, 104  
 Graham, Ronald, 257  
 Grange, Jean-Baptiste, 113FE  
 Graphing calculators  
   button for standard deviation, 549  
   calculating statistics, 4, 65  
   checking Nearly Normal Condition, 542  
   chi-square tests of homogeneity, 638  
   comparing groups with boxplots, 86  
   creating confidence interval, 567–568, 597  
   creating Normal probability plots, 125  
   creating scatterplots, 149–150  
   finding Binomial probabilities, 396  
   finding confidence intervals, 448–449, 510, 541  
   finding correlation, 155  
   finding critical values, 536  
   finding geometric probabilities, 392  
   finding mean of random variables, 370–371  
   finding Normal cutpoints, 119  
   finding Normal percentages, 117–118  
   finding standard deviation of random variables, 370–371  
   finding *t*-model probabilities, 535–536  
   generating random numbers, 262  
   goodness-of-fit test, 624–625  
   inference for regression, 664–665  
   making histograms, 46  
   re-expressing data to achieve linearity, 232  
   regression lines, 187–188  
   residuals plots, 187–188  
   shortcuts to avoid, 234–236  
   straightening curves, 160  
   testing a hypothesis, 468–469, 515, 545–546, 572, 594  
   using logarithmic re-expressions, 233–234  
   working with data, 14–15  
 Groups  
   bimodal distribution and, 52  
   calculator tips for, 86  
   comparing, 82–84, 84FE–85FE  
   comparing means for, 28–2–28–3  
   comparing with boxplots, 83–84, 86  
   comparing with histograms, 82  
   control, 301  
   differences in treatment, 299–300  
   equalizing spread across, 91  
   Independent Groups Assumption, 506, 563–564, 576  
   shifting residuals for, 202–203

## H

Handwashing methods, 28–1–28–2  
 Harvard School of Public Health, 354  
 Harvard University, 427  
 Harvell, Drew, 439  
 HDTV performance, 9FE, 11FE  
 Hepatitis C and tattoos, 633  
 Histograms, 44  
   bimodal, 50  
   calculator tips for, 46  
   common problems, 65  
   comparing groups with, 82  
   describing, 51FE  
   for displaying quantitative data, 44–46  
   gaps in, 45  
   multimodal, 50  
   Nearly Normal Condition and, 537, 542, 550  
   re-expressing data, 224  
   relative frequency, 45  
   sifting residuals for groups, 202  
   skewed, 60  
   symmetric, 50  
   uniform, 50  
   unimodal, 50  
   wind speed example, 82  
 Homogeneity test (chi-square), 627, 629FE–630FE, 638  
 Hopkins Memorial Forest, 80, 82  
 Hurricanes, 146–148, 177FE  
 Hypotheses, 460  
   alternative. *See* Alternative hypothesis  
   null. *See* Null hypothesis  
   writing, 463FE, 481FE  
 Hypothesis testing, 459–479, 480–503  
   calculator tips for, 468–469, 515, 572, 594  
   with chi-square statistic, 621, 622FE–623FE, 624  
   common problems, 474, 496–497  
   on the computer, 476, 498–499  
   confidence intervals and, 487–488, 488FE–490FE, 547  
   effect size and, 492  
   for means, 533, 547  
   Normal model and, 459–461  
   one-sample *t*-test for the mean, 542–543  
   P-value in, 461–462, 465FE, 469–470  
   with paired data, 594  
   power of, 492–494, 493FE, 496FE  
   reasoning of, 463–465  
   sampling variability, 467FE–468FE  
   selecting sex of baby example, 471FE–473FE  
   on snoring, 511  
   standard of reasonable certainty, 462–463  
   Student's *t*-models and, 547  
   threshold value notation, 486  
   trials as, 461  
   Type I error, 491–492, 492FE, 494–496  
   Type II error, 491–492, 494–496  
 Ice breakup times, 661, 662FE–664FE, 664  
 Identifier variables, 12  
 Independence, 29, 326, 349  
   10% Condition, 391  
   Bernoulli trials and, 390–391  
   checking for, 349FE  
   chi-square test for, 632–633, 633FE–636FE, 639  
   common problems, 335, 452, 550  
   conditional probability for, 349

depending on, 350–351  
 disjoint vs., 350  
 DWI test example, 352FE–353FE  
 Independent Groups Assumption, 506, 563–564, 576  
   Multiplication Rule, 331, 332FE  
   of variables, 29, 373–374, 381  
 Independence Assumption  
   for ANOVA, 28–13  
   for Central Limit Theorem, 422  
   in comparing counts, 619–620  
   for comparing means, 563  
   comparing proportions, 506  
   for confidence intervals, 446  
   for multiple regression, 29–5  
   Multiplication Rule and, 331  
   for paired data, 589–590  
   for regression, 652  
   sampling distribution models, 415  
   for Student's *t*-models, 536  
 Independent Groups Assumption, 506, 563–564, 576  
 Independent samples *t*-test, 564  
 Independent variables, 149n  
 Infant mortality, 29–12–29–15  
 Influential points, 206, 206–207  
 Intercept, 176–177, 659  
 International System of Units, 10  
 Internet, data on, 14  
 Interquartile range (IQR), 54–56, 55, 81, 108, 28–14  
 Intersection symbol, 330

## J

Jastrow, J., 296

## K

Kantor, W. M., 257  
 Keno (game), 327  
 Kentucky Derby, 49  
 Keynes, John Maynard, 329  
 Klaussen, Cindy, 587  
 Klüft, Carolina, 104–106  
 Kohavi, Ronny, 8  
 Kostelić, Ivica, 107FE

## L

Ladder of Powers, 226–228, 231  
 Landon, Alf, 269, 284  
 Laplace, Pierre-Simon, 413, 421–422  
 Large Enough Sample Condition, 422–423  
 Law of Averages, 326–327  
 Law of Large Numbers, 326, 421  
 Least significant difference (LSD), 28–20  
 Least Squares method, 172, 179, 649–651, 29–1  
 Left skewness, 51  
 Legionnaires' disease, 293  
 Legitimate probability assignment, 331  
 Level of factor, 294  
 Leverage in linear regression, 206, 206–207  
 Ligety, Ted, 107FE  
 Line of best fit, 172, 175–176  
 Linear association  
   common problems, 160  
   correlation and, 152, 156  
   correlation coefficient and, 161  
   in scatterplots, 147

Linear model, 172, 180–181, 201–202  
 Linear regression, 171–200, 201–221, 649–692.  
*See also Regression*  
 assumptions and conditions, 184–185, 651–653  
 Burger King example, 171–172  
 calculating coefficients, 178FE–180FE  
 causation and, 208–209  
 checking reasonableness, 188  
 common problems, 189, 211–212, 669, 29-18  
 on computers, 192, 213, 672  
 correlation and the line, 173–174  
 extrapolation, 203–205  
 fast food example, 185FE–187FE  
 hurricane example, 177FE  
 hurricane's residual, 180FE  
 influential points in, 206n  
 least squares line, 172–175  
 leverage in, 206–207  
 lurking variables and causation, 208–209  
 Outlier Condition, 184  
 outliers in, 205–206  
 predicted value sizes, 174  
 $R^2$ , 182–184, 657, 29-2  
 residual standard deviation, 181–182, 657  
 residuals in, 180–181, 201–202, 651–652, 654  
 sifting residuals for groups, 202–203  
 subsets in, 203  
 summary values in, 209  
 variation in residuals, 182–183  
 working with multiple methods, 210FE–211FE  
 Linearity Assumption, 184, 201, 651, 29-5  
*Literary Digest*, 269, 284  
 Logarithms, 91, 227, 227n, 233–234, 659FE  
 Lottery, 328  
 Lower quartile, 54, 56  
 Lurking variables, 157, 208, 208–209, 307–308

## M

Margin of error, 443  
 for Bonferroni multiple comparisons, 28-19  
 common problems, 452  
 in confidence intervals, 442–443  
 for difference in independent means, 562  
 for difference in proportions, 507  
 finding, 444FE, 445FE  
 for a mean, 533  
 for a multiple regression coefficient, 29-10  
 polls and, 443FE  
 for a proportion, 531  
 for a regression coefficient, 660  
 Marginal distribution, 24, 26FE  
 Matching, 304  
 in paired data, 588, 598  
 in prospective studies, 304  
 in retrospective studies, 304  
 samples to populations, 270  
 subjects, 304  
 Mean(s), 59, 560–586, 28-1–28-40. *See also*  
 Center of distributions; Expected value  
 assumptions and conditions for, 536–537, 563–564, 564FE  
 calculator tips for, 370–371, 541, 545–546  
 cautions about, 542  
 Central Limit Theorem and, 531, 531FE–532FE  
 common problems, 549–551, 576  
 confidence interval for, 534, 534FE, 538FE–540FE, 541  
 Equal Variance Assumption for, 574  
 grand, in ANOVA model, 28-9  
 and hypothesis tests, 547  
 median compared to, 58–60  
 one-sample  $t$ -interval for the mean, 534, 534FE, 538FE–540FE  
 one-sample  $t$ -test for the mean, 542–543, 543FE–545FE  
 outliers and, 58–60  
 of paired differences, 587–608  
 pooled  $t$ -tests, 574–575  
 of predicted values in regression, 667, 667FE  
 of random variables, 370–371, 372–374, 376FE–377FE, 377, 390  
 sample size and, 547–548  
 sampling distribution models for, 420–421, 423–424, 425FE–426FE  
 and scaling data, 109  
 standard deviation and, 173, 423  
 Student's  $t$ -models, 533, 536–537, 537FE–538FE  
 symmetric distributions and, 58–60  
 testing hypothesis about, 545–546  
 two-sample  $t$ -interval for the difference between means, 562, 564, 565FE–567FE  
 two-sample  $t$ -test for the difference between means, 569–570, 570FE–572FE  
 Median, 53. *See also* Center of distributions  
 of 5-number summary, 56  
 less variable than data, 212  
 resistant, 59  
 Meir, Jessica, 201  
 Metadata, 9n  
 Minimum significant difference (MSD), 28-20  
 M&M's example, 333FE–335FE  
 Mode(s), 49  
 Model(s), 172. *See also* Binomial probability model; Geometric probability model; Linear model; Normal model; Probability models; Sampling distribution models; Student's  $t$ -models  
 ANOVA, 28-9–28-12  
 chi-square, 621, 631, 639  
 conditions for fitting, 203  
 looking beyond data, 89  
 null hypothesis as, 464  
 parameters in, 112  
 for patterns, 80  
 population model, 272  
 random model for simulation, 257  
 usefulness of, 112n  
 Moore, David, 297n  
 Motor vehicle accidents, 354, 530  
 Motorcycle accidents, 480  
 Multimodal distribution, 50, 550  
 Multiple comparisons, 28-19, 28-25  
 Multiple regression, 29-1, 29-1–29-27  
 adjusted  $R^2$ , 29-16  
 ANOVA tables and, 29-9–29-10  
 assumptions and conditions, 29-5–29-7  
 body fat measurement, 29-1, 29-7FE–29-9FE  
 coefficients, 29-3–29-4  
 common problems, 29-17–29-18  
 comparing multiple models, 29-15–29-16  
 on the computer, 29-21  
 Does the Plot Thicken? Condition, 29-6  
 Equal Variance Assumption, 29-6  
 functionality, 29-2–29-4  
 Independence Assumption, 29-5  
 infant mortality, 29-12–29-15  
 Linearity Assumption, 29-5  
 Nearly Normal Condition, 29-6–29-7, 29-13  
 Normality Assumption, 29-6–29-7  
 partial regression plot, 29-3–29-4

Randomization Condition, 29-5  
 sifting residuals for groups, 203n  
 Straight Enough Condition, 29-5  
 testing coefficients, 29-10–29-11  
 Multiplication Rule, 331  
 applying, 332FE  
 General Multiplication Rule, 348, 355–356  
 Multistage sample, 276, 276FE  
 Mutually exclusive events, *See* Disjoint events

## N

Nader, Ralph, 205  
 National Geophysical Data Center (NGDC), 44–45  
 National Highway Traffic Safety Administration, 480, 488FE–490FE, 504, 530  
 National Hurricane Center (NHC), 146–147  
 National Institutes of Health, 108, 294  
 National Sleep Foundation, 511  
 Nearly Normal Condition  
 for ANOVA, 28-15  
 common problems, 550  
 for comparing means, 563  
 histograms and, 542, 550  
 for multiple regression, 29-6–29-7, 29-13  
 Normal models and, 112, 114, 126  
 for paired data, 590  
 for regression, 652  
 for Student's  $t$ -models, 537  
*New England Journal of Medicine*, 481FE, 484FE, 485FE, 492FE  
 NHANES survey, 107–108  
 Nightingale, Florence, 21  
 NOAA (National Oceanic and Atmospheric Administration), 146  
 Nonresponse bias, 283, 283–284  
 Normal model(s), 112  
 68-95-99.7 Rule, 113  
 Binomial models and, 398  
 calculator tips for, 119  
 Central Limit Theorem and, 428  
 common problems, 399  
 critical values from, 534  
 finding percentiles, 116–118  
 hypothesis testing and, 459–461  
 Nearly Normal Condition, 112, 114, 126  
 Normal probability plots, 124–125, 129  
 probability and, 329  
 rules for, 114–116  
 sampling variability and, 415  
 sketching Normal curves, 114  
 standard, 112  
 standard error and, 532  
 Success/Failure Condition, 397  
 working with, 118FE–119FE, 120FE–123FE  
 z-scores and, 111–112, 119  
 Normal percentiles, 116, 116–118, 119  
 Normal Population Assumption  
 for ANOVA, 28-15  
 for comparing means, 563  
 inferences about means, 537  
 for paired data, 590  
 for regression, 652–653  
 Normal probability plots, 124  
 calculator tips for, 125  
 on computers, 129, 552  
 how constructed, 124–125  
 Nearly Normal Condition and, 537  
 Normal probability tables  
 critical values, 445  
 finding Normal percentiles, 116–117  
 Normal scores, 125

Normality Assumption, 112  
 for ANOVA, 28-15  
 common problems, 380  
 inference about means, 549-550  
 for inference in regression, 653  
 for means, 563  
 for multiple regression, 29-6-29-7  
 for paired data, 590  
 Student's *t*-model, 537

Null hypothesis, 460  
 accepting is not possible, 486, 626  
 ANOVA, 28-2  
 chi-square tests, 621, 622n, 626-640  
 choosing, 480-481  
 common problems, 474  
 conclusion about, 465, 465FE  
 conditional probability, 464  
 for difference in proportions, 511-512  
 for goodness-of-fit test, 626  
 in hypothesis testing, 463  
 innocence as the null hypothesis, 462-463  
 multiple regression and, 29-10-29-11  
 one-sample *t*-test, 543  
 P-values and, 461-462  
 paired *t*-test, 591  
 regression, 659-660, 29-2  
 rejecting, 463, 486-487  
 two-sample *t*-test, 569

**O**

Observational studies, 292  
 ANOVA on, 28-21  
 common problems, 430, 28-24  
 designing, 293FE  
 uses for, 293

One-proportion *z*-interval, 442  
 One-proportion *z*-test, 464, 482FE-483FE, 618  
 One-sample *t*-interval for the mean, 534, 534FE, 538FE-540FE  
 One-sample *t*-test for the mean, 542, 542-543, 543FE-545FE  
 One-sided (one-tailed) alternative hypothesis, 466, 485, 669  
 One-way ANOVA *F*-test, 28-15  
 Open Society Institute, 427  
 Ordinary Least Squares, 29-2  
 Outcomes, 325  
 in disjoint events, 331  
 equally likely, 255, 327-328  
 probability of, 342  
 of trials, 258

Outlier Condition  
 for correlation, 153  
 for linear regression, 178, 184  
 for regression, 652

Outliers, 51, 81, 148, 204  
 in ANOVA, 28-13  
 in boxplots, 81, 561  
 checking, 88FE  
 correlation and, 153, 156, 158, 161  
 data analysis of, 87-88  
 in distributions, 87-88  
 far, 81  
 Outlier Condition, 153, 178, 184, 652  
 in paired *t*, 599  
 prefer median to mean, 59-60  
 problems, 59, 126, 550, 599, 669, 28-24, 29-18  
 in regression, 205-206, 654, 29-3-29-4  
 reporting, 88  
 rule of thumb for identifying, 81  
 in scatterplots, 148  
 and standard deviation, 126  
 in Student's *t*, 537, 550  
 wind speed example, 87-88

Overestimate, 172

**P**

P-value, 462  
 as conditional probability, 483-484  
 finding, 465FE  
 high, 484-485  
 in hypothesis testing, 461-462  
 hypothesis testing and, 469-470  
 interpreting, 484FE, 485FE

Paired data, 588  
 10% Condition, 590  
 assumptions and conditions, 589-590, 590FE  
 blocking and, 588, 598  
 calculator tips for, 594  
 common problems, 576, 599  
 differences in means of, 589  
 effect size for, 597, 598FE  
 hypothesis testing with, 594  
 identifying, 588FE  
 Independence Assumption, 589-590  
 Nearly Normal Condition, 590  
 Normal Population Assumption, 590  
 Paired Data Assumption, 589  
 paired-*t* confidence interval, 594-595, 595FE-596FE  
 paired *t*-test, 589, 591, 591FE-593FE  
 Randomization Condition, 589-590

Paired Data Assumption, 589  
 Paired-*t* confidence interval, 594-595, 595FE-596FE  
 Paired *t*-test, 589  
 on the computer, 601-602  
 miles driven by workers, 593FE  
 for paired data, 591  
 speedskater example, 591FE-593FE

Parameters, 112, 272-273, *See also* Model(s)

Partial regression plot, 29-4  
 Participants, 9, 294, 301-302, 304  
 Peirce, C. S., 296, 301n  
 Percentages, 22, 24  
 Percentiles, 56, 116, 116-117  
 Personal probability, 328-329, 329  
 Pew Charitable Trusts, 427  
 Pew Research Center, 268, 271, 399  
 Pie charts, 23, 23-24, 27  
 Pilot study, 281, 284, 309, 450, 548  
 Placebo, 302-303, 303, 462, 480, 512, 28-18  
 Placebo effect, 303  
 Polling methods, 271, 443FE  
 Ponganis, Paul, 201  
 Pooled *t*-intervals, 575  
 Pooled *t* methods, 574-576  
 Pooled *t*-tests, 574, 574-575  
 Pooling, 512, 574  
 in ANOVA, 28-5, 28-15  
 pooled *t*-intervals, 575  
 pooled *t*-tests, 574-575  
 of regression residuals, 669, 29-18  
 two-proportion *z*-test, 511-512

Population(s), 9, 268  
 determining for samples, 270-271, 279-280  
 experiments and random samples, 300-301  
 finite, 391  
 matching samples to, 270  
 parameters, 272-273, 300, 452  
 representative samples from, 9, 270FE

Population parameters, 272  
 common problems, 452  
 sample surveys and, 272-273, 300

Power of hypothesis test, 492-494, 493, 493FE, 496FE

Predicted values, 172  
 confidence intervals for, 667, 667FE  
 size considerations, 174  
 standard errors for, 665-667

Prediction interval for an individual, 667, 667FE

Predictor variable, 149  
 Preusser Group, 480

Probability, 324-341, 326, 342-365  
 Addition Rule, 330-331, 331FE  
 calculator tips for, 535-536  
 common problems, 335  
 Complement Rule, 330, 330FE  
 conditional. *See* Conditional probability  
 continuous, 329  
 empirical, 326  
 formal, 329-332  
 Independence Assumption, 332  
 Law of Large Numbers, 326  
 legitimate probability assignment, 331  
 M&M's example, 333FE-335FE  
 Multiplication Rule, 331, 332FE  
 Normal models and, 329  
 personal, 328-329  
 Probability Assignment Rule, 330-331  
 rules for working with, 329-332  
 theoretical, 327

Probability Assignment Rule, 330, 330-331  
 Probability models, 366, 388-404  
 binomial. *See* Binomial probability model  
 common problems, 380  
 on the computer, 401  
 geometric. *See* Geometric probability model  
 Normal model. *See* Normal model(s)  
 random variables and, 366, 399

Proportion(s), 22, 439-458, 459-479, 504-522  
 10% Condition, 506  
 Central Limit Theorem for, 412-414  
 common problems, 516-517, 549  
 comparing, 504-522  
 on the computer, 476, 519  
 confidence intervals for, 439-458, 447FE-448FE  
 finding standard error of difference, 506FE  
 hypothesis testing, 459-479  
 margin of error and, 531  
 notation for, 273  
 one-proportion *z*-interval, 442  
 one-proportion *z*-test, 464, 482FE-483FE  
 pooling, 512  
 sample considerations, 504  
 sampling distribution models for, 416-417, 417-419, 507  
 standard deviation of difference, 505-506  
 two-proportion *z*-interval, 508, 508FE-510FE  
 two-proportion *z*-test, 512, 513FE-515FE

Prospective studies, 293, 304  
 Pseudorandom numbers, 256  
 Pythagorean Theorem of Statistics, 374, 505, 603, 668

**Q**

Qualitative variable, 10n. *See also* Categorical data

Quantitative data, 44-79  
 5-number summary, 56  
 center of distributions, 49, 52-54  
 common problems, 65-68  
 data analysis considerations, 49  
 displaying on computers, 71  
 dotplots, 49  
 histograms, 44-46  
 sampling distribution models, 419  
 shapes of distributions, 49-52  
 stem-and-leaf displays, 47-48  
 summarizing, 62-63, 63FE-64FE  
 symmetric distributions, 58-60  
 valid surveys and, 280  
 variation in, 62



- Quantitative Data Condition, 49, 651  
 Quantitative variables, 10  
   distribution of, 44  
   linear association between, 152  
   scatterplots for, 147
- Quantitative Variables Condition, 152, 153FE, 178, 184
- Quartiles, 54  
   5-number summary, 56  
   finding, 54, 55  
   lower, 54, 56  
   upper, 54, 56
- Questionnaires, 280–281
- R**
- $R^2$ , 182  
   adjusted, 29–16  
   interpreting, 183FE  
   linear regression and, 29–2  
   not a measure of straightness, 236  
   and  $s_e$ , 657  
   size considerations, 183–184  
   variation in residuals, 182–183
- Random assignment, 294, 297
- Random numbers, 256  
   calculator tips for, 262  
   generating, 256–257  
   to get an SRS, 274FE
- Random phenomenon, 324, 324–326, 330
- Random sampling, *See* Sample(s)
- Random variables, 366, 366–387  
   adding a constant, 372FE–373FE  
   calculator tips for, 370–371  
   common problems, 380–381  
   computers for, 383  
   continuous, 366, 377  
   discrete, 366  
   expected value of, 366–368, 370FE–371FE, 377, 390  
   means and, 370–371, 372–374, 376FE–377FE, 390  
   packaging stereos example, 378FE–380FE  
   probability model, 366  
   Pythagorean Theorem of Statistics, 374  
   restaurant discount example, 368FE  
   standard deviation of, 369, 370–371, 370FE–371FE  
   sum of independent, 373–374, 374FE  
   variance of, 369, 372–374, 376FE–377FE, 424, 505
- Randomization  
   and Central Limit Theorem, 421–422  
   and confidence intervals, 446  
   in experiments, 294, 296  
   in hypothesis testing, 550–551, 563  
   for sample surveys, 270  
   in simulation, 257–259
- Randomization Condition  
   for ANOVA, 28–13  
   Central Limit Theorem, 422, 425FE  
   for chi-square tests, 620, 628  
   in comparing counts, 620  
   for comparing means, 563  
   comparing proportions, 506  
   for confidence intervals, 446  
   for multiple regression, 29–5  
   for paired data, 589–590  
   for regression, 652  
   sampling distribution models, 415  
   for Student's  $t$ -models, 536
- Randomized block design, 304
- Randomness, 255, 255–267  
   building simulations, 258–259  
   card shuffling, 257  
   generating random numbers, 256–257, 262  
   meaning of, 255  
   practical, 257  
   random phenomena, 324–326  
   simulating dice games, 259FE–260FE  
   simulation example, 259FE–260FE  
   simulations on computer, 264
- Range, 54
- Re-expressing data, 90, 222–244  
   calculator tips for, 232, 233–234  
   common problems, 236–237  
   comparing re-expressions, 228FE–230FE, 231FE  
   on computers, 239  
   equalizing spread across groups, 91, 224–225, 28–14  
   equalizing spread across scatterplots, 226  
   goals of, 224–226  
   to improve symmetry, 89–91  
   Ladder of Powers, 226–228, 228FE  
   log-log method, 233  
   recognizing uses, 226FE  
   residuals in, 222–224  
   to straighten curved relationships, 201–202, 210FE–211FE, 222–225, 228FE–230FE, 233, 651, 29–6  
   for symmetry, 89–91, 224, 550  
   Tour de France example, 222
- Regression, 171–200, 201–221, 649–692  
   assumptions and conditions, 184–185, 651–653, 653FE–654FE  
   calculator tips for, 664–665  
   common problems, 669  
   on computers, 192, 213, 672  
   conditions and residuals, 654  
   confidence intervals for predicted values, 667, 667FE  
   and correlation, 173–174  
   Does the Plot Thicken? Condition, 181, 652  
   Equal Variance Assumption, 652  
   extrapolation, 203–205  
   fast food example, 185FE–187FE  
   groups, 202–203  
   ice breakup guess, 661, 662FE–664FE, 664  
   Independence Assumption, 652  
   inferences for, 655FE–656FE, 656–658, 659–660, 664–665  
   influential points in, 205–207  
   intercept in, 176–177, 659–660  
   interpreting model, 660FE  
   least squares criterion, 172–173  
   leverage, 206–207  
   linear model, 172, 650–651  
   Linearity Assumption, 201, 651  
   lurking variables and, 208–209  
   multiple. *See* Multiple regression  
   Nearly Normal Condition, 652  
   Normal Population Assumption, 652–653  
   Outlier Condition, 652  
   outliers, 205–206  
   population and sample, 650–651  
    $R^2$ , 182–184, 183FE  
   Randomization Condition, 652  
   re-expressing to straighten, 202  
   residual standard deviation,  $s_e$ , 181–182, 657  
   residuals, 172, 180–181, 654  
   sampling distribution model for intercept, 658  
   sampling distribution model for slope, 658  
   standard error for predicted values, 665–667  
   standard error for the slope, 658  
   Straight Enough Condition, 651  
   summary variables in, 209  
    $t$ -statistic for slope, 660
- Regression lines, 174, 187–188
- Regression to the mean, 174
- Relative frequency, 22, 45, 326, 347
- Relative frequency bar chart, 23
- Relative frequency histogram, 45
- Relative frequency table, 22
- Replication of experiments, 296
- Representative, 9, 269, 270FE, 273
- Rescaling data, 108, 108–109, 109FE
- Research hypothesis, 570n. *See also* Alternative Hypothesis
- Residual(s), 172  
   in ANOVA, 28–5  
   for chi-square, 619, 623, 631, 632FE, 636–637  
   groups in, 202–203  
   hurricane example, 180FE  
   influential points in, 207  
   least squares, 172, 29–1  
   linear models and, 180–181, 201–202, 651, 29–1  
   in re-expressing data, 201, 222–224  
   standard deviation of, 181–182  
   standardized, 631, 632FE  
   variation in, 182–183
- Residual standard deviation  $s_e$ , 181–182, 657, 28–12
- Residuals plots, 181, 187–188
- Resistant, median as, 59
- Respondents, 9, 268, 271, 333
- Response bias, 282, 283–284
- Response variables, 149, 294  
   determining, 295FE  
   in experiments, 294  
   in simulations, 258–259
- Retrospective studies, 292, 304
- Reverse conditioning, 356, 357FE–358FE, 359FE
- Rho ( $\rho$ ) for correlation, 273
- Richter scale, 44–45, 44n
- Right skewness, 51, 60
- Roosevelt, Franklin Delano, 269, 284
- Rounding, 67, 548n
- S**
- (residual standard deviation), 181–182, 657, 28–12
- Sample(s), 9, 268–291, 269  
   bias, 269, 274, 282FE, 283FE  
   cluster, 275–276  
   common problems, 451, 550–551, 640  
   on the computer, 287  
   confidence interval for, 490  
   convenience, 282–283  
   determining populations, 279–280  
   experiments and, 300–301  
   Large Enough Sample Condition, 422–423  
   matching to populations, 270  
   multistage, 276, 276FE  
   paired data, 587–598  
   random, 270  
   regression and, 650–651  
   representative, 270FE, 273  
   response, 282  
   Simple Random Sample, 273, 274FE  
   stratified, 274, 275FE  
   systematic, 277  
   voluntary response, 282  
   watching TV example, 277FE–279FE
- Sample size  
   choosing, 449–450, 450FE  
   finding, 548FE–549FE  
   heart attack risk example, 496FE  
   means and, 547–548  
   regression inference and, 658  
   Sample Size Assumption, 415, 422, 446, 620

- Sample Size Assumption  
 Central Limit Theorem, 422  
 for chi-square, 620  
 for proportions, 446  
 sampling distribution models, 415
- Sample space, 325, 342
- Sample statistic, 272, 480
- Sample surveys, 268–291  
 census considerations, 271–272  
 cluster sampling, 275–276  
 common problems, 282–284  
 determining populations, 279–280  
 examining part of the whole, 268–270  
 population parameters, 272–273, 300  
 randomizing, 270  
 sample size for, 270–271  
 sampling example, 277FE–279FE  
 Simple Random Sample, 273  
 stratified random sampling, 274  
 systematic samples, 277  
 valid, 280–281
- Sampling distribution models, 412–438, 413  
 10% Condition, 415, 422  
*aspergillosis* example, 439  
 assumptions and conditions, 415–416  
 Central Limit Theorem, 412–414, 421–422, 429, 532  
 common problems, 429–430  
 confidence intervals, 440–442  
 for difference between means, 563  
 for difference between proportions, 507  
 hypothesis testing and, 459–461  
 Independence Assumption, 415  
 for a mean, 420–421, 423–424, 425FE–426FE  
 Normal model and, 415, 429  
 for a proportion, 416–417, 417–419  
 Randomization Condition, 415, 422  
 for regression slopes, 658  
 Sample Size Assumption, 415  
 Success/Failure Condition, 415–416  
 summarized, 429  
 variation in, 427
- Sampling error, 274, 414
- Sampling frame, 273, 280, 283
- Sampling variability, 274, 414, 467FE–468FE
- SAT tests, 110FE–111FE, 114FE–115FE, 116–117, 118FE–120FE
- Scales  
 combining data on different scales, 105  
*f*/stop, 222  
 measurement, 10  
 no effect on correlation, 156, 156FE  
 Richter, 44–45, 44n
- Scatterplot matrix, 29–12–29–13
- Scatterplots, 146–170, 147  
 association, 147–150  
 axes, 149  
 calculator tips for, 149–150  
 common problems, 160–161  
 on computers, 163  
 curved patterns in, 234  
 direction, 147  
 emperor penguins example, 202  
 form of, 147  
 hurricane example, 146–148, 148FE, 153FE  
 outliers in, 148  
 for quantitative variables, 147  
 re-expressing data, 225–226  
 of residuals, 181, 187–188  
 roles for variables, 148–150  
 standardizing, 151  
 straightening, 158–159, 228FE–230FE, 231FE  
 strength, 148  
 summary values in, 209  
 variables in, 148–150
- Sea fans, 439
- Seat-belt use, 504, 508FE–510FE
- Segmented bar charts, 30, 30–33
- Shapes of distributions, 49  
 flight cancellation example, 56FE–58FE  
 gaps in, 52  
 modes of histograms, 49  
 outliers, 51  
 standardizing z-scores, 110  
 symmetric histograms, 50
- Shifting data, 107–108, 108
- Significance level, 486, 486–487
- Similar Spread Condition, 574, 28–14
- Simple Random Sample (SRS), 273, 274FE
- Simpson's paradox, 34–35, 35
- Simulation(s), 255–267, 257  
 building, 258–259  
 common problems, 263  
 components of, 258  
 on computers, 264  
 of dice games, 259FE–260FE  
 lottery for dorm room example, 260FE–262FE  
 response variables in, 258–259  
 sampling distributions of a mean, 420–421  
 trials and, 258
- Simulation component, 258
- Single-blind experiments, 302
- Skewed distributions, 50  
 common problems, 430, 550  
 re-expressing to improve symmetry, 89–91, 224  
 Student's *t*-models and, 537
- Skewed left, 60
- Skewed right, 60
- Skujyte, Austr, 104–106
- Slope, 176–177  
 inference for, 659–660  
 and influential points, 207  
 interpreting, 176–177, 177FE, 649  
 parameter ( $\beta_1$ ), 272
- Slope-intercept form, 668
- Smoking, cancer and, 157–158
- Something has to Happen Rule, *See*  
 Probability Assignment Rule
- Speed skaters, 587
- SPLOM, 29–12–29–13
- Spread, 49, 54  
 comparing, 82, 28–2  
 Does the Plot Thicken? Condition, 181, 652, 29–6  
 equalizing across groups, 91, 224–225  
 interquartile range, 54–56  
 range, 54  
 regression inference and, 657  
 of residuals, 657
- Similar Spread Condition, 574, 28–14
- standard deviation, 60–61, 104n  
 standardizing, 110
- SRS (Simple Random Sample), 273, 274FE
- Stacked format for data, 28–27
- Standard deviation(s), 60–61, 369  
 calculator tips for, 370–371, 546  
 Central Limit Theorem and, 531  
 common problems, 126  
 of difference between means, 561, 562FE  
 of difference between proportions, 505–506  
 for discrete random variables, 369, 370FE–371FE  
 finding, 61  
 of the mean, 423  
 Normal models, rules of, 114–116  
 Normal models, working with, 120FE–123FE  
 Normal models and z-scores, 111–112  
 of a proportion, 413  
 of a random variable, 370–371  
 rescaling data, 108–109  
 of residuals, 181–182, 657, 28–12–28–13  
 as ruler, 104–105  
 spread, 60–61
- standardized variables, 110FE–111FE  
 testing hypothesis about a mean, 546  
 z-scores, 105–107, 110, 111–112, 119
- Standard error(s), 440  
 calculating, 668  
 comparing means, 561–563  
 of difference between means, 561–562  
 of difference between proportions, 505–506, 511  
 of a mean, 532  
 Normal model and, 532  
 for paired difference, 591  
 for predicted values, 665–667  
 for proportion, 440, 487n  
 for regression slope, 652, 658
- Standard Normal distribution, 112
- Standard Normal model, 112
- Standardized residuals for chi-square, 631, 632FE
- Standardized values, 105
- Standardized variables, 110FE–111FE
- Standardizing  
 plotting standardized values to find  
 correlation, 151  
 skiing times, 106FE  
 standard deviations and, 105  
 z-scores, 110
- Statistic, 112, 272, 480, 620–621
- Statistical significance, 299, 486  
 common problems, 497  
 for means, 546  
 practical significance vs., 487  
 in treatment group differences, 299–300
- Statistics, 2, 2–6, 112  
 Binomial probability model for Bernoulli trials, 394  
 calculator tips for, 4, 65  
 finding mean of random variables, 390  
 Florence Nightingale and, 21  
 line of best fit, 175–176  
 Normal model considerations, 398  
 notation in, 58  
 standard deviation as a ruler, 104–105  
 statistical inference and, 483n
- Stem-and-leaf displays, 47, 47–48
- Stemplot, 47–48
- Straight Enough Condition, 152  
 for correlation, 161  
 for multiple regression, 29–5  
 for regression, 178, 201, 202, 224, 231, 651
- Strata, 274
- Stratified random sample, 274, 275FE
- Student's *t*-models, 533  
 10% Condition, 536–537  
 assumptions and conditions, 536–537, 537FE–538FE  
 calculator tips for, 535–536  
 critical value from, 534, 540  
 degrees of freedom and, 533, 535  
 Gosset and, 533  
 hypothesis testing and, 547  
 Independence Assumption, 536  
 Nearly Normal Condition, 537, 550  
 paired-*t* confidence interval, 595FE–596FE  
 Randomization Condition, 536  
 standard error and, 533  
 two-sample *t* methods, 579
- Subjects, 9, 294, 301–302, 304
- Subsets in regression, 203
- Success/Failure Condition  
 for Binomial models, 397  
 comparing proportions, 512–513  
 confidence interval for small samples, 490  
 for proportions, 415–416, 446, 460, 507
- Symmetric distributions, 50  
 in Student's *t*-models, 536–537  
 summarizing, 58–60
- Systematic sample, 277

**T**

*t*-ratios for regression coefficients, 29-10–29-11, 29-17

*t*-tests

- one-sample, for mean, 542–543
- paired, 591
- pooled, 574–575
- for regression slope, 29-10
- two-sample, for means, 569

Tables

- ANOVA, 28-7–28-8
- cells of, 24, 619
- conditional probability and, 351
- contingency. *See* Contingency tables
- frequency, 21, 21–22
- organizing values, 8–9
- two-way, 619, 633

Tails, of distribution, 50

Tchebycheff, Pafnuty, 116

Theoretical probability, 327

Therapeutic touch, 484–485

TI Tips. *See* Graphing calculators

Timeplots, 88–89

*Titanic* example, 20

Transforming data. *See* Re-expressing data

Treatment(s), 294

- assessing effect of, 300
- blinding subjects to, 301–302
- determining, 295FE
- diagrams for, 297
- differences in groups, 299–300
- randomization of, 294, 296

Treatment Mean Square ( $MS_T$ ), 28-5

- degrees of freedom and, 28-6
- handwashing example, 28-7

Treatment Sum of Squares, 28-11

Tree diagrams, 354, 354–356

Trials, 258, 325, 461. *See also* Bernoulli trials

Tsunamis, 44, 52–54

Tukey, John W., 47, 82, 441

Two-factor experiments, completely randomized, 305

Two-proportion *z*-interval, 508

- finding, 508FE–510FE
- seat belt use example, 508FE–510FE

Two-proportion *z*-test, 513

- online safety example, 516FE
- snoring example, 513FE–515FE

Two-sample *t*-interval for the difference

- between means, 562, 564, 565FE–567FE

Two-sample *t* methods, 562

- common problems, 599
- on the computer, 579
- two-sample *t*-interval for the difference
  - between means, 562, 564, 565FE–567FE
- two-sample *t*-test for the difference
  - between means, 562, 569–570, 570FE–572FE

Two-sample *t*-test for the difference

- between means, 562, 569, 569–570, 570FE–572FE

Two-sided (two-tailed) alternative hypothesis, 466

Two-way tables, 619, 633

Type I error, 491, 491–492

- effect size and, 494–495
- heart attack risk example, 492FE, 496FE
- reducing, 485–486

Type II error, 491, 491–492

- effect size and, 494–495
- heart attack risk example, 496FE
- reducing, 485–486

**U**

Undercounting population, 272

Undercoverage, 283

Underestimate, 172

Uniform distribution, 50

Unimodal distribution, 50, 535–537. *See also* Nearly Normal Condition

Union symbol, 330

Units, 10, 156

Upper quartile, 54, 56

U.S. Geological Survey, 44n

**V**

Vague concepts, 52, 156

Variables, 9

- associations between, 29FE
- blocking, 296
- categorical, 10, 10–12, 22
- causal relationships, 157
- dependent, 149n
- distributions of, 224
- explanatory, 149, 294
- identifier, 12
- independence of, 29, 373–374, 381
- independent, 149n
- lurking, 157, 208–209, 307–308
- predictor, 149
- quantitative, 10, 44
- random. *See* Random variables
- response, 149, 258–259, 294, 295FE
- in scatterplots, 148–150
- skewed, 60
- standardized, 110FE–111FE

Variance, 61, 369

- addition rule for. *See* Pythagorean Theorem of Statistics
- Equal Variance Assumption, 574, 652, 28-14, 28-20, 29-6

- of independent random variables, 381, 424, 505
- of random variables, 369, 372–374, 376FE–377FE

## Variation

- controlling sources of, 295
- in quantitative data, 62
- in residuals, 182–183
- in sampling distribution models, 427

Venn, John, 329

Venn diagrams

- contingency tables and, 351
- creation of, 329
- food survey example, 344FE
- General Addition Rule, 343

Voluntary response bias, 282

Voluntary response sample, 282

**W**

Wainer, Howard, 427

Wind speed, 80

Within Mean Square, 28-5

Women's Health Initiative, 294

Woods, Tiger, 388–389

**X**

*x*-axis, 149

*x*-variable, 149

**Y**

*y*-axis, 149

*y*-intercept, 176–177

*y*-variable, 149

**Z**

*z*-scores, 105, 112

- calculator tips for, 117–118
- combining, 107FE
- Normal models and, 111–112
- Normal percentiles and, 119
- in scatterplots, 152
- standardizing, 105–107, 110

Zabriskie, Dave, 222

Zodiac signs, 618

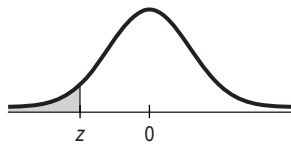
Zwerling, Harris, 427

# Index of TI Tips

- 1-PropZInt, 448, 469
- 1-Var Stats, 65
- 2-PropZInt, 510, 515
- 2-SampTInterval, 545–546
- 2-SampTTest, 568
- add data to a list, 15
- binomial probabilities, 396
- boxplot, 86
- change data, 14
- Chi-square
  - homogeneity, 638
  - independence, 638
  - $\chi^2$  GOF-Test, 624–625
  - $\chi^2$  Test, 638
- clear a data list, 15
- confidence interval for
  - difference of independent means, 545–546
  - difference of two proportions, 510
  - mean of paired differences, 594
  - one mean, 541
  - one proportion, 448–449
  - slope of regression line, 667
- correlation, 155
- curved models, 234–235
- data, 14–15
- delete data from a list, 15
- DiagnosticOn, 155
- edit data, 14
- enter data, 14
- ERR: DIM MISMATCH, 46, 150
- ERR: DOMAIN, 449
- expected value, 371–372
- ExpReg, 234
- five-number summary, 65
- frequency table, 46
- geometric probabilities, 392
- histogram, create a, 46
  - of residuals, 666–667
- hypothesis test for,
  - difference of independent means, 568
  - difference of proportions, 515
  - homogeneity, 638
  - independence, 638
  - mean of paired differences, 594
  - one mean, 545–546
  - one proportion, 469
- insert data into a list, 15
- invNorm, 119
- invT, 535
- LinReg, 155, 187
- LinRegTInt, 667
- list is missing, 15
- LIST NAMES, 150, 155
- LnReg, 235
- logarithms, 232, 233–234
- matrix, 638
- mean of,
  - random variable, 371–372
  - sample data, 65
- median, 65
- naming lists, 150
- Normal model, 112–113
- Normal percentiles, 118, 119
- Normal probability plot, 125
- numerical summary, 65
- PwrReg, 235
- QuadReg, 235
- quartiles, 65
- random numbers, 262
- random variables, 371–372
- re-expressing data, 160, 232, 233–234
- regression line, 187
- residuals, 188
- restore missing data list, 15
- scatterplot, 149–150
- slope of regression line, 667
- standard deviation of,
  - random variable, 371–372
  - sample data, 65
- STAT PLOT 5, 86, 125, 149, 188
- TInterval, 541, 594
- t*-models, 535–536
- TRACE a,
  - boxplot, 86
  - histogram, 46
  - scatterplot, 150
- T-Test, 545–546, 597

# Tables

| Row | TABLE OF RANDOM DIGITS |       |       |       |       |       |       |       |       |       |
|-----|------------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 1   | 96299                  | 07196 | 98642 | 20639 | 23185 | 56282 | 69929 | 14125 | 38872 | 94168 |
| 2   | 71622                  | 35940 | 81807 | 59225 | 18192 | 08710 | 80777 | 84395 | 69563 | 86280 |
| 3   | 03272                  | 41230 | 81739 | 74797 | 70406 | 18564 | 69273 | 72532 | 78340 | 36699 |
| 4   | 46376                  | 58596 | 14365 | 63685 | 56555 | 42974 | 72944 | 96463 | 63533 | 24152 |
| 5   | 47352                  | 42853 | 42903 | 97504 | 56655 | 70355 | 88606 | 61406 | 38757 | 70657 |
| 6   | 20064                  | 04266 | 74017 | 79319 | 70170 | 96572 | 08523 | 56025 | 89077 | 57678 |
| 7   | 73184                  | 95907 | 05179 | 51002 | 83374 | 52297 | 07769 | 99792 | 78365 | 93487 |
| 8   | 72753                  | 36216 | 07230 | 35793 | 71907 | 65571 | 66784 | 25548 | 91861 | 15725 |
| 9   | 03939                  | 30763 | 06138 | 80062 | 02537 | 23561 | 93136 | 61260 | 77935 | 93159 |
| 10  | 75998                  | 37203 | 07959 | 38264 | 78120 | 77525 | 86481 | 54986 | 33042 | 70648 |
| 11  | 94435                  | 97441 | 90998 | 25104 | 49761 | 14967 | 70724 | 67030 | 53887 | 81293 |
| 12  | 04362                  | 40989 | 69167 | 38894 | 00172 | 02999 | 97377 | 33305 | 60782 | 29810 |
| 13  | 89059                  | 43528 | 10547 | 40115 | 82234 | 86902 | 04121 | 83889 | 76208 | 31076 |
| 14  | 87736                  | 04666 | 75145 | 49175 | 76754 | 07884 | 92564 | 80793 | 22573 | 67902 |
| 15  | 76488                  | 88899 | 15860 | 07370 | 13431 | 84041 | 69202 | 18912 | 83173 | 11983 |
| 16  | 36460                  | 53772 | 66634 | 25045 | 79007 | 78518 | 73580 | 14191 | 50353 | 32064 |
| 17  | 13205                  | 69237 | 21820 | 20952 | 16635 | 58867 | 97650 | 82983 | 64865 | 93298 |
| 18  | 51242                  | 12215 | 90739 | 36812 | 00436 | 31609 | 80333 | 96606 | 30430 | 31803 |
| 19  | 67819                  | 00354 | 91439 | 91073 | 49258 | 15992 | 41277 | 75111 | 67496 | 68430 |
| 20  | 09875                  | 08990 | 27656 | 15871 | 23637 | 00952 | 97818 | 64234 | 50199 | 05715 |
| 21  | 18192                  | 95308 | 72975 | 01191 | 29958 | 09275 | 89141 | 19558 | 50524 | 32041 |
| 22  | 02763                  | 33701 | 66188 | 50226 | 35813 | 72951 | 11638 | 01876 | 93664 | 37001 |
| 23  | 13349                  | 46328 | 01856 | 29935 | 80563 | 03742 | 49470 | 67749 | 08578 | 21956 |
| 24  | 69238                  | 92878 | 80067 | 80807 | 45096 | 22936 | 64325 | 19265 | 37755 | 69794 |
| 25  | 92207                  | 63527 | 59398 | 29818 | 24789 | 94309 | 88380 | 57000 | 50171 | 17891 |
| 26  | 66679                  | 99100 | 37072 | 30593 | 29665 | 84286 | 44458 | 60180 | 81451 | 58273 |
| 27  | 31087                  | 42430 | 60322 | 34765 | 15757 | 53300 | 97392 | 98035 | 05228 | 68970 |
| 28  | 84432                  | 04916 | 52949 | 78533 | 31666 | 62350 | 20584 | 56367 | 19701 | 60584 |
| 29  | 72042                  | 12287 | 21081 | 48426 | 44321 | 58765 | 41760 | 43304 | 13399 | 02043 |
| 30  | 94534                  | 73559 | 82135 | 70260 | 87936 | 85162 | 11937 | 18263 | 54138 | 69564 |
| 31  | 63971                  | 97198 | 40974 | 45301 | 60177 | 35604 | 21580 | 68107 | 25184 | 42810 |
| 32  | 11227                  | 58474 | 17272 | 37619 | 69517 | 62964 | 67962 | 34510 | 12607 | 52255 |
| 33  | 28541                  | 02029 | 08068 | 96656 | 17795 | 21484 | 57722 | 76511 | 27849 | 61738 |
| 34  | 11282                  | 43632 | 49531 | 78981 | 81980 | 08530 | 08629 | 32279 | 29478 | 50228 |
| 35  | 42907                  | 15137 | 21918 | 13248 | 39129 | 49559 | 94540 | 24070 | 88151 | 36782 |
| 36  | 47119                  | 76651 | 21732 | 32364 | 58545 | 50277 | 57558 | 30390 | 18771 | 72703 |
| 37  | 11232                  | 99884 | 05087 | 76839 | 65142 | 19994 | 91397 | 29350 | 83852 | 04905 |
| 38  | 64725                  | 06719 | 86262 | 53356 | 57999 | 50193 | 79936 | 97230 | 52073 | 94467 |
| 39  | 77007                  | 26962 | 55466 | 12521 | 48125 | 12280 | 54985 | 26239 | 76044 | 54398 |
| 40  | 18375                  | 19310 | 59796 | 89832 | 59417 | 18553 | 17238 | 05474 | 33259 | 50595 |

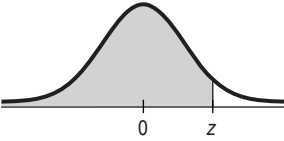


**Table Z**  
Areas under the  
standard normal curve

| Second decimal place in z |        |        |        |        |        |        |        |        |        | z      |      |
|---------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|------|
| 0.09                      | 0.08   | 0.07   | 0.06   | 0.05   | 0.04   | 0.03   | 0.02   | 0.01   | 0.00   |        |      |
| 0.0001                    | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | -3.8 |
| 0.0001                    | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | -3.7 |
| 0.0001                    | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0002 | 0.0002 | -3.6 |
| 0.0002                    | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | 0.0002 | -3.5 |
| 0.0002                    | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | -3.4 |
| 0.0003                    | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0004 | 0.0005 | 0.0005 | 0.0005 | -3.3 |
| 0.0005                    | 0.0005 | 0.0005 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0006 | 0.0007 | 0.0007 | -3.2 |
| 0.0007                    | 0.0007 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0010 | -3.1 |
| 0.0010                    | 0.0010 | 0.0011 | 0.0011 | 0.0011 | 0.0012 | 0.0012 | 0.0012 | 0.0013 | 0.0013 | 0.0013 | -3.0 |
| 0.0014                    | 0.0014 | 0.0015 | 0.0015 | 0.0016 | 0.0016 | 0.0017 | 0.0018 | 0.0018 | 0.0018 | 0.0019 | -2.9 |
| 0.0019                    | 0.0020 | 0.0021 | 0.0021 | 0.0022 | 0.0023 | 0.0023 | 0.0024 | 0.0025 | 0.0025 | 0.0026 | -2.8 |
| 0.0026                    | 0.0027 | 0.0028 | 0.0029 | 0.0030 | 0.0031 | 0.0032 | 0.0033 | 0.0034 | 0.0034 | 0.0035 | -2.7 |
| 0.0036                    | 0.0037 | 0.0038 | 0.0039 | 0.0040 | 0.0041 | 0.0043 | 0.0044 | 0.0045 | 0.0045 | 0.0047 | -2.6 |
| 0.0048                    | 0.0049 | 0.0051 | 0.0052 | 0.0054 | 0.0055 | 0.0057 | 0.0059 | 0.0060 | 0.0060 | 0.0062 | -2.5 |
| 0.0064                    | 0.0066 | 0.0068 | 0.0069 | 0.0071 | 0.0073 | 0.0075 | 0.0078 | 0.0080 | 0.0080 | 0.0082 | -2.4 |
| 0.0084                    | 0.0087 | 0.0089 | 0.0091 | 0.0094 | 0.0096 | 0.0099 | 0.0102 | 0.0104 | 0.0104 | 0.0107 | -2.3 |
| 0.0110                    | 0.0113 | 0.0116 | 0.0119 | 0.0122 | 0.0125 | 0.0129 | 0.0132 | 0.0136 | 0.0136 | 0.0139 | -2.2 |
| 0.0143                    | 0.0146 | 0.0150 | 0.0154 | 0.0158 | 0.0162 | 0.0166 | 0.0170 | 0.0174 | 0.0174 | 0.0179 | -2.1 |
| 0.0183                    | 0.0188 | 0.0192 | 0.0197 | 0.0202 | 0.0207 | 0.0212 | 0.0217 | 0.0222 | 0.0222 | 0.0228 | -2.0 |
| 0.0233                    | 0.0239 | 0.0244 | 0.0250 | 0.0256 | 0.0262 | 0.0268 | 0.0274 | 0.0281 | 0.0281 | 0.0287 | -1.9 |
| 0.0294                    | 0.0301 | 0.0307 | 0.0314 | 0.0322 | 0.0329 | 0.0336 | 0.0344 | 0.0351 | 0.0351 | 0.0359 | -1.8 |
| 0.0367                    | 0.0375 | 0.0384 | 0.0392 | 0.0401 | 0.0409 | 0.0418 | 0.0427 | 0.0436 | 0.0436 | 0.0446 | -1.7 |
| 0.0455                    | 0.0465 | 0.0475 | 0.0485 | 0.0495 | 0.0505 | 0.0516 | 0.0526 | 0.0537 | 0.0537 | 0.0548 | -1.6 |
| 0.0559                    | 0.0571 | 0.0582 | 0.0594 | 0.0606 | 0.0618 | 0.0630 | 0.0643 | 0.0655 | 0.0655 | 0.0668 | -1.5 |
| 0.0681                    | 0.0694 | 0.0708 | 0.0721 | 0.0735 | 0.0749 | 0.0764 | 0.0778 | 0.0793 | 0.0793 | 0.0808 | -1.4 |
| 0.0823                    | 0.0838 | 0.0853 | 0.0869 | 0.0885 | 0.0901 | 0.0918 | 0.0934 | 0.0951 | 0.0951 | 0.0968 | -1.3 |
| 0.0985                    | 0.1003 | 0.1020 | 0.1038 | 0.1056 | 0.1075 | 0.1093 | 0.1112 | 0.1131 | 0.1131 | 0.1151 | -1.2 |
| 0.1170                    | 0.1190 | 0.1210 | 0.1230 | 0.1251 | 0.1271 | 0.1292 | 0.1314 | 0.1335 | 0.1335 | 0.1357 | -1.1 |
| 0.1379                    | 0.1401 | 0.1423 | 0.1446 | 0.1469 | 0.1492 | 0.1515 | 0.1539 | 0.1562 | 0.1562 | 0.1587 | -1.0 |
| 0.1611                    | 0.1635 | 0.1660 | 0.1685 | 0.1711 | 0.1736 | 0.1762 | 0.1788 | 0.1814 | 0.1814 | 0.1841 | -0.9 |
| 0.1867                    | 0.1894 | 0.1922 | 0.1949 | 0.1977 | 0.2005 | 0.2033 | 0.2061 | 0.2090 | 0.2090 | 0.2119 | -0.8 |
| 0.2148                    | 0.2177 | 0.2206 | 0.2236 | 0.2266 | 0.2296 | 0.2327 | 0.2358 | 0.2389 | 0.2389 | 0.2420 | -0.7 |
| 0.2451                    | 0.2483 | 0.2514 | 0.2546 | 0.2578 | 0.2611 | 0.2643 | 0.2676 | 0.2709 | 0.2709 | 0.2743 | -0.6 |
| 0.2776                    | 0.2810 | 0.2843 | 0.2877 | 0.2912 | 0.2946 | 0.2981 | 0.3015 | 0.3050 | 0.3050 | 0.3085 | -0.5 |
| 0.3121                    | 0.3156 | 0.3192 | 0.3228 | 0.3264 | 0.3300 | 0.3336 | 0.3372 | 0.3409 | 0.3409 | 0.3446 | -0.4 |
| 0.3483                    | 0.3520 | 0.3557 | 0.3594 | 0.3632 | 0.3669 | 0.3707 | 0.3745 | 0.3783 | 0.3783 | 0.3821 | -0.3 |
| 0.3859                    | 0.3897 | 0.3936 | 0.3974 | 0.4013 | 0.4052 | 0.4090 | 0.4129 | 0.4168 | 0.4168 | 0.4207 | -0.2 |
| 0.4247                    | 0.4286 | 0.4325 | 0.4364 | 0.4404 | 0.4443 | 0.4483 | 0.4522 | 0.4562 | 0.4562 | 0.4602 | -0.1 |
| 0.4641                    | 0.4681 | 0.4721 | 0.4761 | 0.4801 | 0.4840 | 0.4880 | 0.4920 | 0.4960 | 0.4960 | 0.5000 | -0.0 |

For  $z \leq -3.90$ , the areas are 0.0000 to four decimal places.

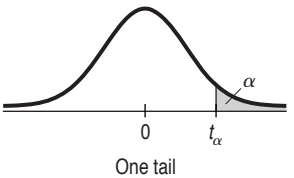
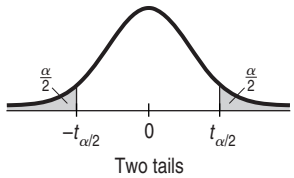
**Table Z (cont.)**  
Areas under the standard normal curve



| <i>z</i> | Second decimal place in <i>z</i> |             |             |             |             |             |             |             |             |             |
|----------|----------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|          | <i>0.00</i>                      | <i>0.01</i> | <i>0.02</i> | <i>0.03</i> | <i>0.04</i> | <i>0.05</i> | <i>0.06</i> | <i>0.07</i> | <i>0.08</i> | <i>0.09</i> |
| 0.0      | 0.5000                           | 0.5040      | 0.5080      | 0.5120      | 0.5160      | 0.5199      | 0.5239      | 0.5279      | 0.5319      | 0.5359      |
| 0.1      | 0.5398                           | 0.5438      | 0.5478      | 0.5517      | 0.5557      | 0.5596      | 0.5636      | 0.5675      | 0.5714      | 0.5753      |
| 0.2      | 0.5793                           | 0.5832      | 0.5871      | 0.5910      | 0.5948      | 0.5987      | 0.6026      | 0.6064      | 0.6103      | 0.6141      |
| 0.3      | 0.6179                           | 0.6217      | 0.6255      | 0.6293      | 0.6331      | 0.6368      | 0.6406      | 0.6443      | 0.6480      | 0.6517      |
| 0.4      | 0.6554                           | 0.6591      | 0.6628      | 0.6664      | 0.6700      | 0.6736      | 0.6772      | 0.6808      | 0.6844      | 0.6879      |
| 0.5      | 0.6915                           | 0.6950      | 0.6985      | 0.7019      | 0.7054      | 0.7088      | 0.7123      | 0.7157      | 0.7190      | 0.7224      |
| 0.6      | 0.7257                           | 0.7291      | 0.7324      | 0.7357      | 0.7389      | 0.7422      | 0.7454      | 0.7486      | 0.7517      | 0.7549      |
| 0.7      | 0.7580                           | 0.7611      | 0.7642      | 0.7673      | 0.7704      | 0.7734      | 0.7764      | 0.7794      | 0.7823      | 0.7852      |
| 0.8      | 0.7881                           | 0.7910      | 0.7939      | 0.7967      | 0.7995      | 0.8023      | 0.8051      | 0.8078      | 0.8106      | 0.8133      |
| 0.9      | 0.8159                           | 0.8186      | 0.8212      | 0.8238      | 0.8264      | 0.8289      | 0.8315      | 0.8340      | 0.8365      | 0.8389      |
| 1.0      | 0.8413                           | 0.8438      | 0.8461      | 0.8485      | 0.8508      | 0.8531      | 0.8554      | 0.8577      | 0.8599      | 0.8621      |
| 1.1      | 0.8643                           | 0.8665      | 0.8686      | 0.8708      | 0.8729      | 0.8749      | 0.8770      | 0.8790      | 0.8810      | 0.8830      |
| 1.2      | 0.8849                           | 0.8869      | 0.8888      | 0.8907      | 0.8925      | 0.8944      | 0.8962      | 0.8980      | 0.8997      | 0.9015      |
| 1.3      | 0.9032                           | 0.9049      | 0.9066      | 0.9082      | 0.9099      | 0.9115      | 0.9131      | 0.9147      | 0.9162      | 0.9177      |
| 1.4      | 0.9192                           | 0.9207      | 0.9222      | 0.9236      | 0.9251      | 0.9265      | 0.9279      | 0.9292      | 0.9306      | 0.9319      |
| 1.5      | 0.9332                           | 0.9345      | 0.9357      | 0.9370      | 0.9382      | 0.9394      | 0.9406      | 0.9418      | 0.9429      | 0.9441      |
| 1.6      | 0.9452                           | 0.9463      | 0.9474      | 0.9484      | 0.9495      | 0.9505      | 0.9515      | 0.9525      | 0.9535      | 0.9545      |
| 1.7      | 0.9554                           | 0.9564      | 0.9573      | 0.9582      | 0.9591      | 0.9599      | 0.9608      | 0.9616      | 0.9625      | 0.9633      |
| 1.8      | 0.9641                           | 0.9649      | 0.9656      | 0.9664      | 0.9671      | 0.9678      | 0.9686      | 0.9693      | 0.9699      | 0.9706      |
| 1.9      | 0.9713                           | 0.9719      | 0.9726      | 0.9732      | 0.9738      | 0.9744      | 0.9750      | 0.9756      | 0.9761      | 0.9767      |
| 2.0      | 0.9772                           | 0.9778      | 0.9783      | 0.9788      | 0.9793      | 0.9798      | 0.9803      | 0.9808      | 0.9812      | 0.9817      |
| 2.1      | 0.9821                           | 0.9826      | 0.9830      | 0.9834      | 0.9838      | 0.9842      | 0.9846      | 0.9850      | 0.9854      | 0.9857      |
| 2.2      | 0.9861                           | 0.9864      | 0.9868      | 0.9871      | 0.9875      | 0.9878      | 0.9881      | 0.9884      | 0.9887      | 0.9890      |
| 2.3      | 0.9893                           | 0.9896      | 0.9898      | 0.9901      | 0.9904      | 0.9906      | 0.9909      | 0.9911      | 0.9913      | 0.9916      |
| 2.4      | 0.9918                           | 0.9920      | 0.9922      | 0.9925      | 0.9927      | 0.9929      | 0.9931      | 0.9932      | 0.9934      | 0.9936      |
| 2.5      | 0.9938                           | 0.9940      | 0.9941      | 0.9943      | 0.9945      | 0.9946      | 0.9948      | 0.9949      | 0.9951      | 0.9952      |
| 2.6      | 0.9953                           | 0.9955      | 0.9956      | 0.9957      | 0.9959      | 0.9960      | 0.9961      | 0.9962      | 0.9963      | 0.9964      |
| 2.7      | 0.9965                           | 0.9966      | 0.9967      | 0.9968      | 0.9969      | 0.9970      | 0.9971      | 0.9972      | 0.9973      | 0.9974      |
| 2.8      | 0.9974                           | 0.9975      | 0.9976      | 0.9977      | 0.9977      | 0.9978      | 0.9979      | 0.9979      | 0.9980      | 0.9981      |
| 2.9      | 0.9981                           | 0.9982      | 0.9982      | 0.9983      | 0.9984      | 0.9984      | 0.9985      | 0.9985      | 0.9986      | 0.9986      |
| 3.0      | 0.9987                           | 0.9987      | 0.9987      | 0.9988      | 0.9988      | 0.9989      | 0.9989      | 0.9989      | 0.9990      | 0.9990      |
| 3.1      | 0.9990                           | 0.9991      | 0.9991      | 0.9991      | 0.9992      | 0.9992      | 0.9992      | 0.9992      | 0.9993      | 0.9993      |
| 3.2      | 0.9993                           | 0.9993      | 0.9994      | 0.9994      | 0.9994      | 0.9994      | 0.9994      | 0.9995      | 0.9995      | 0.9995      |
| 3.3      | 0.9995                           | 0.9995      | 0.9995      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9996      | 0.9997      |
| 3.4      | 0.9997                           | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9997      | 0.9998      |
| 3.5      | 0.9998                           | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      | 0.9998      |
| 3.6      | 0.9998                           | 0.9998      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |
| 3.7      | 0.9999                           | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |
| 3.8      | 0.9999                           | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      | 0.9999      |

For  $z \geq 3.90$ , the areas are 1.0000 to four decimal places.

|                          |          | 0.20  | 0.10  | 0.05   | 0.02   | 0.01   |          |
|--------------------------|----------|-------|-------|--------|--------|--------|----------|
| Two tail probability     |          |       |       |        |        |        |          |
| One tail probability     |          | 0.10  | 0.05  | 0.025  | 0.01   | 0.005  |          |
| <b>Table T</b>           | df       |       |       |        |        |        | df       |
| Values of $t_\alpha$     | 1        | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 1        |
|                          | 2        | 1.886 | 2.920 | 4.303  | 6.965  | 9.925  | 2        |
|                          | 3        | 1.638 | 2.353 | 3.182  | 4.541  | 5.841  | 3        |
|                          | 4        | 1.533 | 2.132 | 2.776  | 3.747  | 4.604  | 4        |
|                          | 5        | 1.476 | 2.015 | 2.571  | 3.365  | 4.032  | 5        |
|                          | 6        | 1.440 | 1.943 | 2.447  | 3.143  | 3.707  | 6        |
|                          | 7        | 1.415 | 1.895 | 2.365  | 2.998  | 3.499  | 7        |
|                          | 8        | 1.397 | 1.860 | 2.306  | 2.896  | 3.355  | 8        |
|                          | 9        | 1.383 | 1.833 | 2.262  | 2.821  | 3.250  | 9        |
|                          | 10       | 1.372 | 1.812 | 2.228  | 2.764  | 3.169  | 10       |
|                          | 11       | 1.363 | 1.796 | 2.201  | 2.718  | 3.106  | 11       |
|                          | 12       | 1.356 | 1.782 | 2.179  | 2.681  | 3.055  | 12       |
|                          | 13       | 1.350 | 1.771 | 2.160  | 2.650  | 3.012  | 13       |
|                          | 14       | 1.345 | 1.761 | 2.145  | 2.624  | 2.977  | 14       |
|                          | 15       | 1.341 | 1.753 | 2.131  | 2.602  | 2.947  | 15       |
|                          | 16       | 1.337 | 1.746 | 2.120  | 2.583  | 2.921  | 16       |
|                          | 17       | 1.333 | 1.740 | 2.110  | 2.567  | 2.898  | 17       |
|                          | 18       | 1.330 | 1.734 | 2.101  | 2.552  | 2.878  | 18       |
|                          | 19       | 1.328 | 1.729 | 2.093  | 2.539  | 2.861  | 19       |
|                          | 20       | 1.325 | 1.725 | 2.086  | 2.528  | 2.845  | 20       |
|                          | 21       | 1.323 | 1.721 | 2.080  | 2.518  | 2.831  | 21       |
|                          | 22       | 1.321 | 1.717 | 2.074  | 2.508  | 2.819  | 22       |
|                          | 23       | 1.319 | 1.714 | 2.069  | 2.500  | 2.807  | 23       |
|                          | 24       | 1.318 | 1.711 | 2.064  | 2.492  | 2.797  | 24       |
|                          | 25       | 1.316 | 1.708 | 2.060  | 2.485  | 2.787  | 25       |
|                          | 26       | 1.315 | 1.706 | 2.056  | 2.479  | 2.779  | 26       |
|                          | 27       | 1.314 | 1.703 | 2.052  | 2.473  | 2.771  | 27       |
|                          | 28       | 1.313 | 1.701 | 2.048  | 2.467  | 2.763  | 28       |
|                          | 29       | 1.311 | 1.699 | 2.045  | 2.462  | 2.756  | 29       |
|                          | 30       | 1.310 | 1.697 | 2.042  | 2.457  | 2.750  | 30       |
|                          | 32       | 1.309 | 1.694 | 2.037  | 2.449  | 2.738  | 32       |
|                          | 35       | 1.306 | 1.690 | 2.030  | 2.438  | 2.725  | 35       |
|                          | 40       | 1.303 | 1.684 | 2.021  | 2.423  | 2.704  | 40       |
|                          | 45       | 1.301 | 1.679 | 2.014  | 2.412  | 2.690  | 45       |
|                          | 50       | 1.299 | 1.676 | 2.009  | 2.403  | 2.678  | 50       |
|                          | 60       | 1.296 | 1.671 | 2.000  | 2.390  | 2.660  | 60       |
|                          | 75       | 1.293 | 1.665 | 1.992  | 2.377  | 2.643  | 75       |
|                          | 100      | 1.290 | 1.660 | 1.984  | 2.364  | 2.626  | 100      |
|                          | 120      | 1.289 | 1.658 | 1.980  | 2.358  | 2.617  | 120      |
|                          | 140      | 1.288 | 1.656 | 1.977  | 2.353  | 2.611  | 140      |
|                          | 180      | 1.286 | 1.653 | 1.973  | 2.347  | 2.603  | 180      |
|                          | 250      | 1.285 | 1.651 | 1.969  | 2.341  | 2.596  | 250      |
|                          | 400      | 1.284 | 1.649 | 1.966  | 2.336  | 2.588  | 400      |
|                          | 1000     | 1.282 | 1.646 | 1.962  | 2.330  | 2.581  | 1000     |
|                          | $\infty$ | 1.282 | 1.645 | 1.960  | 2.326  | 2.576  | $\infty$ |
| <b>Confidence levels</b> |          | 80%   | 90%   | 95%    | 98%    | 99%    |          |





| Right tail probability         |     | 0.10    | 0.05    | 0.025   | 0.01    | 0.005   |
|--------------------------------|-----|---------|---------|---------|---------|---------|
| <b>Table <math>\chi</math></b> | df  |         |         |         |         |         |
| Values of $\chi^2_{\alpha}$    | 1   | 2.706   | 3.841   | 5.024   | 6.635   | 7.879   |
|                                | 2   | 4.605   | 5.991   | 7.378   | 9.210   | 10.597  |
|                                | 3   | 6.251   | 7.815   | 9.348   | 11.345  | 12.838  |
|                                | 4   | 7.779   | 9.488   | 11.143  | 13.277  | 14.860  |
|                                | 5   | 9.236   | 11.070  | 12.833  | 15.086  | 16.750  |
|                                | 6   | 10.645  | 12.592  | 14.449  | 16.812  | 18.548  |
|                                | 7   | 12.017  | 14.067  | 16.013  | 18.475  | 20.278  |
|                                | 8   | 13.362  | 15.507  | 17.535  | 20.090  | 21.955  |
|                                | 9   | 14.684  | 16.919  | 19.023  | 21.666  | 23.589  |
|                                | 10  | 15.987  | 18.307  | 20.483  | 23.209  | 25.188  |
|                                | 11  | 17.275  | 19.675  | 21.920  | 24.725  | 26.757  |
|                                | 12  | 18.549  | 21.026  | 23.337  | 26.217  | 28.300  |
|                                | 13  | 19.812  | 22.362  | 24.736  | 27.688  | 29.819  |
|                                | 14  | 21.064  | 23.685  | 26.119  | 29.141  | 31.319  |
|                                | 15  | 22.307  | 24.996  | 27.488  | 30.578  | 32.801  |
|                                | 16  | 23.542  | 26.296  | 28.845  | 32.000  | 34.267  |
|                                | 17  | 24.769  | 27.587  | 30.191  | 33.409  | 35.718  |
|                                | 18  | 25.989  | 28.869  | 31.526  | 34.805  | 37.156  |
|                                | 19  | 27.204  | 30.143  | 32.852  | 36.191  | 38.582  |
|                                | 20  | 28.412  | 31.410  | 34.170  | 37.566  | 39.997  |
|                                | 21  | 29.615  | 32.671  | 35.479  | 38.932  | 41.401  |
|                                | 22  | 30.813  | 33.924  | 36.781  | 40.290  | 42.796  |
|                                | 23  | 32.007  | 35.172  | 38.076  | 41.638  | 44.181  |
|                                | 24  | 33.196  | 36.415  | 39.364  | 42.980  | 45.559  |
|                                | 25  | 34.382  | 37.653  | 40.647  | 44.314  | 46.928  |
|                                | 26  | 35.563  | 38.885  | 41.923  | 45.642  | 48.290  |
|                                | 27  | 36.741  | 40.113  | 43.195  | 46.963  | 49.645  |
|                                | 28  | 37.916  | 41.337  | 44.461  | 48.278  | 50.994  |
|                                | 29  | 39.087  | 42.557  | 45.722  | 49.588  | 52.336  |
|                                | 30  | 40.256  | 43.773  | 46.979  | 50.892  | 53.672  |
|                                | 40  | 51.805  | 55.759  | 59.342  | 63.691  | 66.767  |
|                                | 50  | 63.167  | 67.505  | 71.420  | 76.154  | 79.490  |
|                                | 60  | 74.397  | 79.082  | 83.298  | 88.381  | 91.955  |
|                                | 70  | 85.527  | 90.531  | 95.023  | 100.424 | 104.213 |
|                                | 80  | 96.578  | 101.879 | 106.628 | 112.328 | 116.320 |
|                                | 90  | 107.565 | 113.145 | 118.135 | 124.115 | 128.296 |
|                                | 100 | 118.499 | 124.343 | 129.563 | 135.811 | 140.177 |

