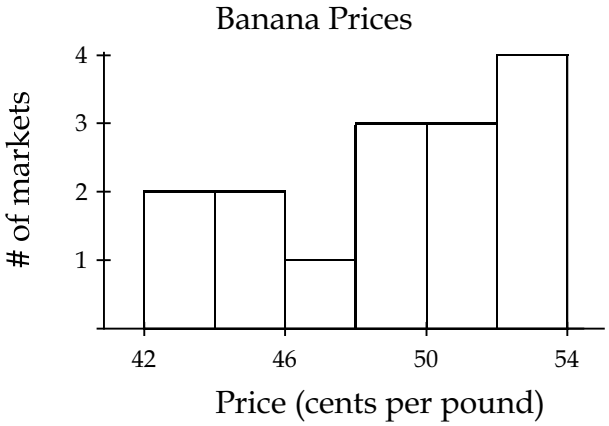
Review of Part I – Exploring and Understanding Data

1. Bananas.



a) A histogram of the prices of bananas from 15 markets, as reported by the USDA, appears at the right.

b) The distribution of banana prices is skewed to the left, so median and IQR are appropriate measures of center and spread.
Median = 49 cents per pound
IQR = 7 cents per pound

c) The distribution of the prices of bananas from 15 markets, as reported by the USDA, is unimodal and skewed to the left. The center of the distribution is approximately 50 cents, with the lowest price 42 cents per pound and the highest price 53 cents per pound.

2. Prenatal care.

a) $\frac{5.4+3.9+6.1}{3} = 5.1\overline{3}$, so the overall rate of 5.1 deaths per thousand live births is equal to the average of the rates for Intensive, Adequate, and Inadequate prenatal care, when rounded to the nearest tenth. There is no reason this should be the case unless the number of women receiving each type of prenatal care is approximately the same.

b) Yes, the results indicate (but do not prove) that adequate prenatal care is important for pregnant women. The mortality rate is quite a bit lower for women with adequate care than for other women.

c) No, the results do not suggest that a woman pregnant with twins should be wary of seeking too much medical care. Intensive care is given for emergency conditions. The data do not suggest that the level of care is the cause of the higher mortality.

3. Singers.

a) The two statistics could be the same if there were many sopranos of that height.

b) The distribution of heights of each voice part is roughly symmetric. The basses and tenors are generally taller than the altos and sopranos, with the basses being slightly taller than the tenors. The sopranos and altos have about the same median height. Heights of basses and sopranos are more consistent than altos and tenors.

4. Dialysis.

There are only three patients currently on dialysis. With so few patients, no display is needed. We know that one patient has had his or her toes amputated and that two patients have developed blindness. What we don't know is whether or not the patient that has had his or her toes amputated has also developed blindness. Even if we wanted to, we do not have enough information to make an appropriate display.

5. Beanstalks.

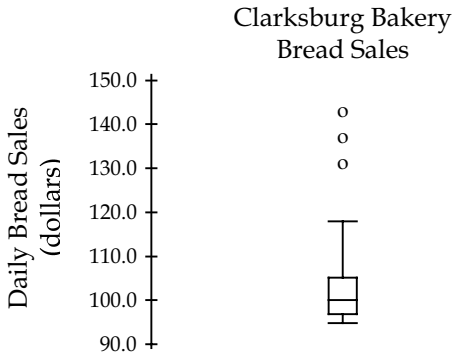
a) The greater standard deviation for the distribution of women's heights means that their heights are more variable than the heights of men.

b) The z-score for women to qualify is 2.4 compared with 1.75 for men, so it is harder for women to qualify.

6. Bread.

a) The distribution of the number of loaves sold each day in the last 100 days at the Clarksburg Bakery is unimodal and skewed to the right. The mode is near 100, with the majority of days recording fewer than 120 loaves sold. The number of loaves sold ranges from 95 to 140.

b) The mean number of loaves sold will be higher than the median number of loaves sold, since the distribution of sales is skewed to the right. The mean is sensitive to this skewness, while the median is resistant.



c) Create a boxplot with quartiles at 97 and 105.5, median at 100. The IQR is 8.5 so the upper fence is at 105.5 + 1.5(8.5) = 118.25. There are several high outliers. There are no low outliers because the min at 95 lies well within the lower fence at 97 – 1.5(8.5) = 84.25. One possible boxplot is at the right.

d) The distribution of daily bread sales is not symmetric, but rather skewed to the right. The Normal model is not appropriate for this distribution. No conclusions can be drawn.

7. State University.

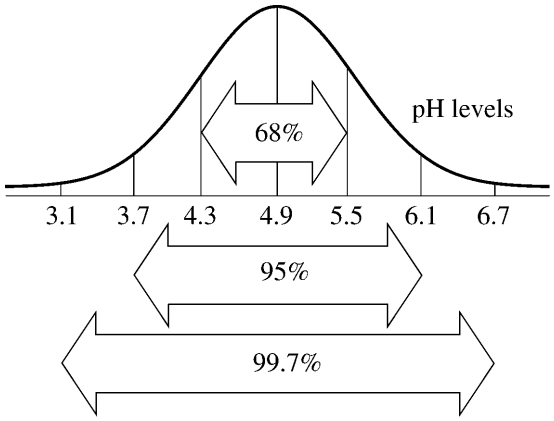
a) *Who* – Local residents near State University. *What* – Age, whether or not the respondent attended college, and whether or not the respondent had a favorable opinion of State University. *When* – Not specified. *Where* – Region around State University. *Why* – The information will be included in a report to the University's directors. *How* – 850 local residents were surveyed by phone.

b) There is one quantitative variable, age, probably measured in years. There are two categorical variables, college attendance (yes or no), and opinion of State University (favorable or unfavorable).

c) There are several problems with the design of the survey. No mention is made of a random selection of residents. Furthermore, there may be a non-response bias present. People with an unfavorable opinion of the university may hang up as soon as the staff member identifies himself or herself. Also, response bias may be introduced by the interviewer. The responses of the residents may be influenced by the fact that employees of the university are asking the questions. There may be greater percentage of favorable responses to the survey than truly exist.

8. Acid Rain.

a) The Normal model for pH level of rainfall in the Shenandoah Mountains is at the right.

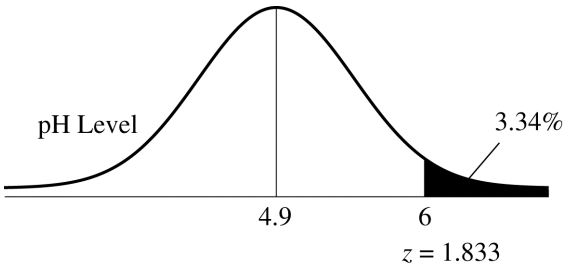


b)

$$z = \frac{y - \mu}{\sigma}$$

$$z = \frac{6 - 4.9}{0.6}$$

$$z = 1.833$$



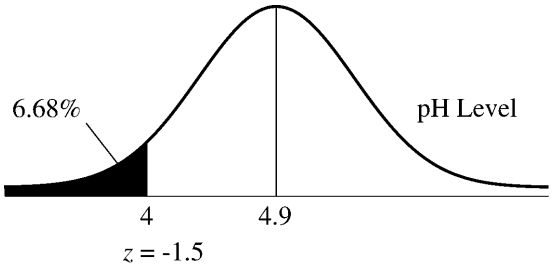
According to the Normal model, 3.34% of the rainstorms are expected to produce rainfall with pH levels above 6.

c)

$$z = \frac{y - \mu}{\sigma}$$

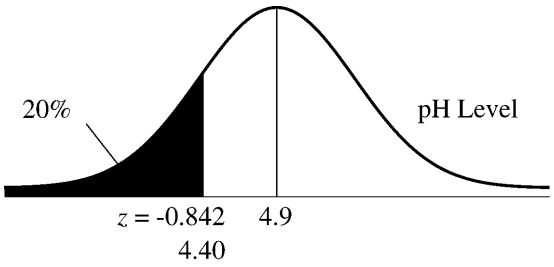
$$z = \frac{4 - 4.9}{0.6}$$

$$z = -1.5$$



According to the Normal model, 6.68% of rainstorms are expected to produce rainfall with pH levels below 4.

d)



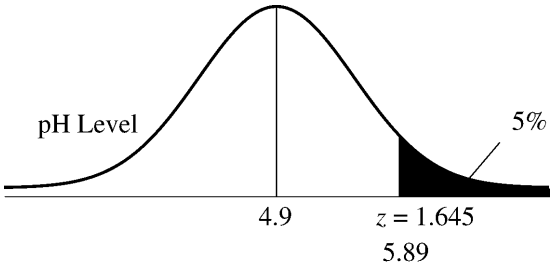
$$z = \frac{y - \mu}{\sigma}$$

$$-0.842 = \frac{y - 4.9}{0.6}$$

$$y = 4.40$$

According to the Normal model, the most acidic 20% of storms have pH below 4.40.

e)



$$z = \frac{y - \mu}{\sigma}$$

$$1.645 = \frac{y - 4.9}{0.6}$$

$$y = 5.89$$

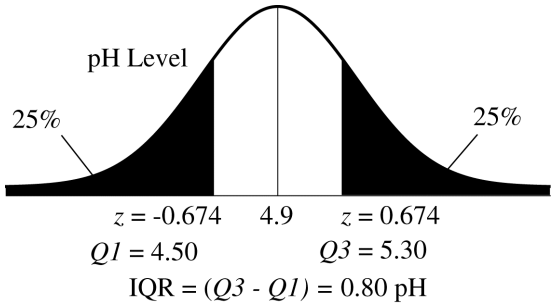
According to the Normal model, the least acidic 5% of storms have pH above 5.89.

f)

$$z = \frac{y - \mu}{\sigma}$$

$$-0.674 = \frac{Q1 - 4.9}{0.6}$$

$$Q1 = 4.50$$



$$z = \frac{y - \mu}{\sigma}$$

$$0.674 = \frac{Q3 - 4.9}{0.6}$$

$$Q3 = 5.30$$

According to the Normal model, the IQR of the pH levels of the rainstorms is 0.80.

9. Fraud detection.

a) Even though they are numbers, the SIC code is a categorical variable. A histogram is a quantitative display, so it is not appropriate.

b) The Normal model will not work at all. The Normal model is for modeling distributions of unimodal and symmetric quantitative variables. SIC code is a categorical variable.

10. Streams.

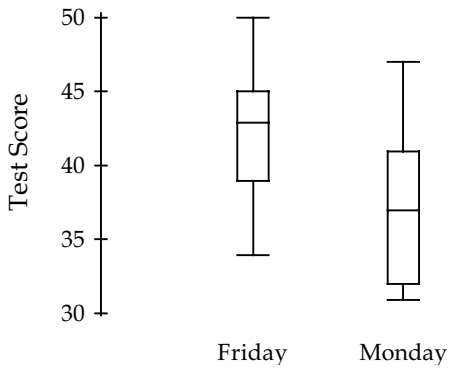
a) Stream Name – categorical; Substrate – categorical; pH – quantitative; Temperature – quantitative (^oC); BCI – quantitative.

b) Substrate is a categorical variable, so a pie chart or a bar chart would be a useful display.

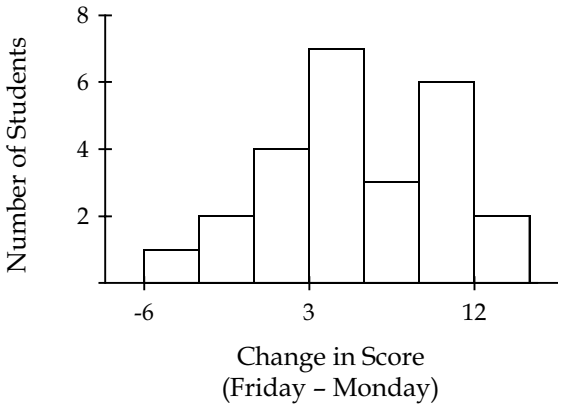
11. Cramming.

a) Comparitive boxplots of the distributions of Friday and Monday scores are at the right.

b) The distribution of scores on Friday was generally higher by about 5 points. Students fared worse on Monday after preparing for the test on Friday. The spreads are about the same, but the scores on Monday are slightly skewed to the right.



c) A histogram of the distribution of change in test score is at the right.



d) The distribution of changes in score is roughly unimodal and symmetric, and is centered near 4 points. Changes ranged from a student who scored 5 points higher on Monday, to two students who each scored 14 points higher on Friday. Only three students did better on Monday.

12. Computers and Internet.

The conclusion is not sound. Many homes may have both a personal computer and access to the Internet. (In Chapter 14, we will say that these percentages may not be added because they are not disjoint.)

13. Let's play cards.

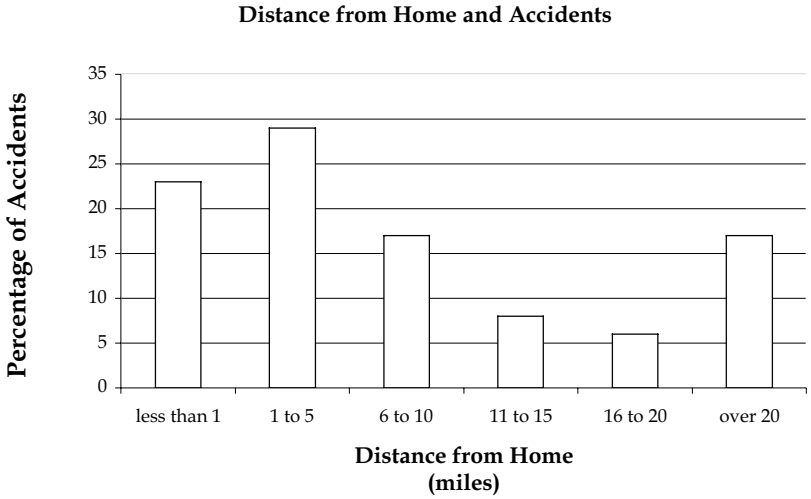
a) Suit is a categorical variable.

b) In the game of Go Fish, the denomination is not ordered. Numbers are merely matched with one another. You may have seen children's Go Fish decks that have symbols or pictures on the cards instead of numbers. These work just fine.

c) In the game of Gin Rummy, the order of the cards is important. During the game, ordered "runs" of cards are assembled (with Jack = 11, Queen = 12, King = 13), and at the end of the hand, points are totaled from the denomination of the card (face cards = 10 points). However, even in Gin Rummy, the denomination of the card sometimes behaves like a categorical variable. When you are collecting 3s, for example, order doesn't matter.

14. Accidents.

a) The distances from home are organized in categories, so a bar chart is provided at the right. A pie chart would also be useful, since the percentages represent parts of a whole.



b) A greater percentage of accidents happen close to home than further away. But it is likely that people drive more miles close to home as well. These data do not indicate that driving near home is dangerous.

15. Hard water.

a) The variables in this study are both quantitative. Annual mortality rate for males is measured in deaths per 100,000. Calcium concentration is measured in parts per million.

b) The distribution of calcium concentration is skewed right, with many towns having concentrations below 25 ppm. The rest of the towns have calcium concentrations which are distributed in a fairly uniform pattern from 25 ppm to 100 ppm, tapering off to a maximum concentration around 150 ppm.
The distribution of mortality rates is unimodal and symmetric, with center approximately 1500 deaths per 100,000. The distribution has a range of 1000 deaths per 100,000, from 1000 to 2000 deaths per 100,000.

16. Hard water II.

a) The overall mean mortality rate is $\frac{34(1631.59)+27(1388.85)}{34+27}=1524.15$ deaths per 100,000.

b) The distribution of mortality rates for the towns north of Derby is generally higher than the distribution of mortality rates for the towns south of Derby. Fully half of the towns south of Derby have mortality rates lower than any of the towns north of Derby. A quarter of the northern towns have rates higher than any of the Southern towns.

17. Seasons.

a) The two histograms have different horizontal and vertical scales. This makes a quick comparison impossible.

b) The center of the distribution of average temperatures in January in is the low 30s, compared to a center of the distribution of July temperatures in the low 70s. The January distribution is also much more spread out than the July distribution. The range is over 50 degrees in January, compared to a range of over 20 in July. The distribution of average temperature in January is skewed slightly to the right, while the distribution of average temperature in July is roughly symmetric.

c) The distribution of difference in average temperature (July – January) for 60 large U.S. cities is slightly skewed to the left, with median at approximately 44 degrees. There are several low outliers, cities with very little difference between their average July and January temperatures. The single high outlier is a city with a large difference in average temperature between July and January. The middle 50% of differences are between approximately 38 and 46 degrees.

18. Old Faithful.

The distribution of time gaps between eruptions of Old Faithful is bimodal. A large cluster of time gaps has a mode at approximately 80 minutes and a slightly smaller cluster of time gaps has a mode at approximately 50 minutes. The distribution around each mode is fairly symmetric.

19. Old Faithful?

a) The distribution of duration of the 222 eruptions is bimodal, with modes at approximately 2 minutes and 4.5 minutes. The distribution is fairly symmetric around each mode.

b) The bimodal shape of the distribution of duration of the 222 eruptions suggests that there may be two distinct groups of eruption durations. Summary statistics would try to summarize these two groups as a single group, which wouldn't make sense.

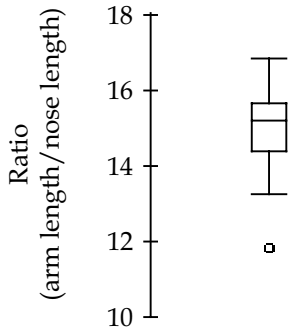
c) The intervals between eruptions are generally longer for long eruptions than the intervals for short eruptions. Over 75% of the short eruptions had intervals of approximately 60 minutes or less, while almost all of the long eruptions had intervals of more than 60 minutes.

20. Teen drivers.

Involvement in fatal crashes is not independent of age. If the variables were independent, we would expect the percentage of fatal crashes involving teen drivers to be the same as the overall percentage of teen drivers.

21. Liberty's nose.

a) The distribution of the ratio of arm length to nose length of 18 girls in a statistics class is unimodal and roughly symmetric, with center around 15. There is one low outlier, a ratio of 11.8. A boxplot is provided at the right. A histogram or stemplot is also an appropriate display.

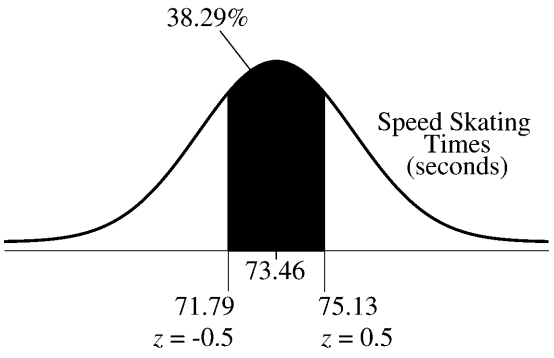


b) In the presence of an outlier, the 5-number summary is the appropriate choice for summary statistics. The 5-number summary is 11.8, 14.4, 15.25, 15.7, 16.9. The IQR is 1.3.

c) The ratio of 9.3 for the Statue of Liberty is very low, well below the lowest ratio in the statistics class, 11.8, which is already a low outlier. Compared to the girls in the statistics class, the Statue of Liberty's nose is very long in relation to her arm.

22. Winter Olympics 2002 speed skating.

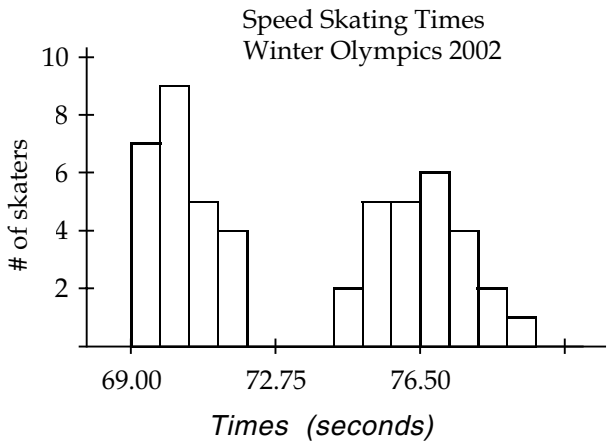
a)



Being within 1.67 seconds of the mean when the standard deviation is 3.33 seconds is the same as being within half a standard deviation of the mean. If the Normal model is appropriate, we expect approximately 38.29% of the speed skating times to be within this interval.

b) Only three speed skating times, 71.96, 74.75, and 74.94 seconds were within the interval 71.79 – 75.13 seconds. This is only 6% (3 of 50) of the times.

c) A histogram of the distribution of speed skating times is at the right. The distribution is bimodal, reflecting the male and female times clustered around each mode. There were two distinct groups of times within this distribution, and it probably should have been displayed as two distributions. At any rate, the Normal model is not appropriate, since the distribution of the speed skating times is not unimodal and symmetric.



23. Sample.

Overall, the follow-up group was insured only 11.1% of the time as compared to 16.6% for the not traced group. At first, it appears that group is associated with presence of health insurance. But for blacks, the follow-up group was quite close (actually slightly higher) in terms of being insured: 8.9% to 8.7%. The same is true for whites. The follow-up group was insured 83.3% of the time, compared to 82.5% of the not traced group. When broken down by race, we see that group is not associated with presence of health insurance for either race. This demonstrates Simpson's paradox, because the overall percentages lead us to believe that there is an association between health insurance and group, but we see the truth when we examine the situation more carefully.

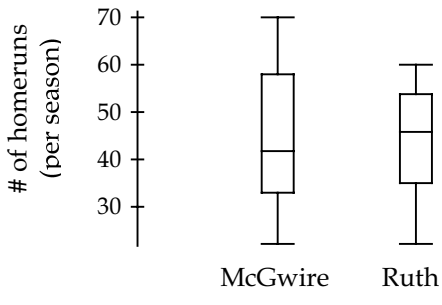
24. Sluggers.

a) The 5-number summary for McGwire's career is 3, 25.5, 36, 50.5, 70. The IQR is 25.

b) By the outlier test, 1.5(IQR) = 37.5. There are no homerun totals less than Q1 – 37.5 or greater than Q3 + 37.5. Technically, there are no outliers. However, the seasons in which McGwire hit fewer than 22 homeruns stand out as a separate group.

c) Parallel boxplots comparing the homerun careers of Mark McGwire and Babe Ruth are at the right.

d) Without the injured seasons, McGwire and Ruth's home run production distributions look similar. (Note: Ruth's seasons as a pitcher were not included.) Ruth's median is a little higher, and he was a little more consistent (less spread), but McGwire had the two highest season totals.



e) A back-to-back stem-and-leaf display of the homerun careers of McGwire and Ruth is at the right.

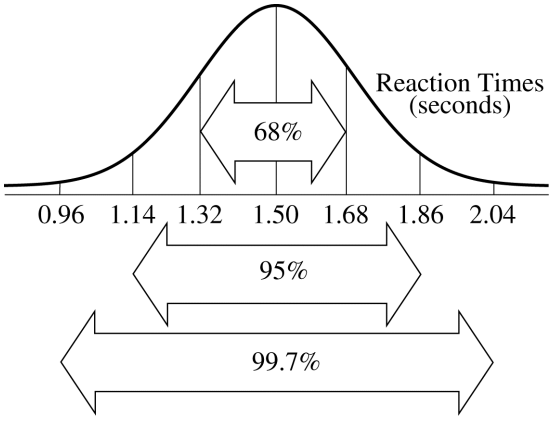
Ruth		McGwire
	7	0
0	6	5
944	5	28
9766611	4	29
54	3	2399
52	2	2

7|0 = 70 homeruns per season

f) From the stem-and-leaf display, we can see that Ruth was much more consistent. During most of his seasons, Ruth had homerun totals in the 40s and 50s. The shape of McGwire's distribution of homeruns is revealed to be skewed to the right.

25. Be quick!

a) The Normal model for the distribution of reaction times is at the right.



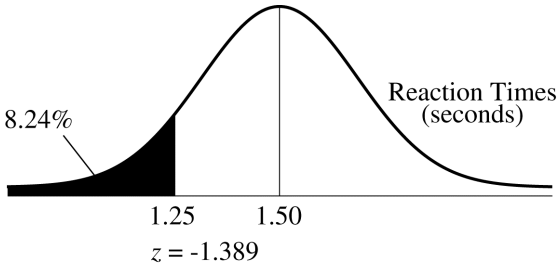
b) The distribution of reaction times is unimodal and symmetric, with mean 1.5 seconds, and standard deviation 0.18 seconds. According to the Normal model, 95% of drivers are expected to have reaction times between 1.14 seconds and 1.86 seconds.

c)

$$z = \frac{y - \mu}{\sigma}$$

$$z = \frac{1.25 - 1.50}{0.18}$$

$$z = -1.389$$



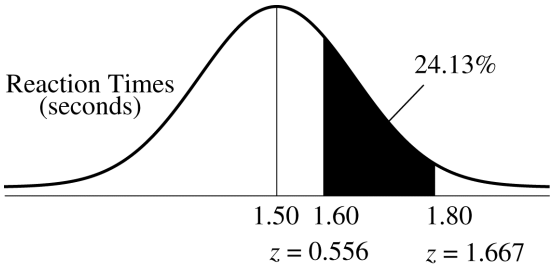
According to the Normal model, 8.24% of drivers are expected to have reaction times below 1.25 seconds.

d)

$$z = \frac{y - \mu}{\sigma}$$

$$z = \frac{1.6 - 1.5}{0.18}$$

$$z = 0.556$$



$$z = \frac{y - \mu}{\sigma}$$

$$z = \frac{1.8 - 1.5}{0.18}$$

$$z = 1.667$$

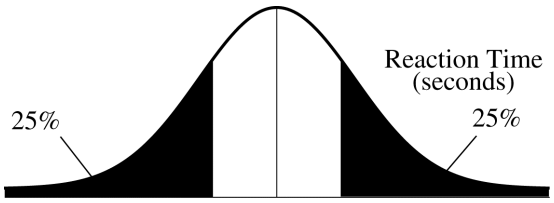
According to the Normal model, 24.13% of drivers are expected to have reaction times between 1.6 seconds and 1.8 seconds.

e)

$$z = \frac{y - \mu.}{\sigma}$$

$$-0.674 = \frac{Q1 - 1.5}{0.18}$$

$$Q1 = 1.38$$



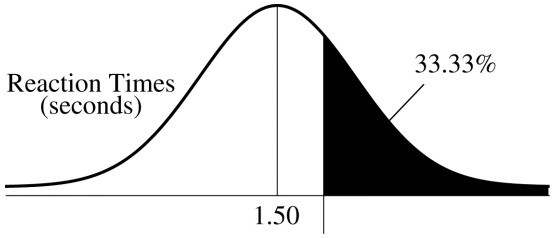
$$z = \frac{y - \mu}{\sigma}$$

$$0.674 = \frac{Q3 - 1.5}{0.18}$$

$$Q3 = 1.62$$

According to the Normal model, the interquartile range of the distribution of reaction times is expected to be 0.24 seconds.

f)



$$z = \frac{y - \mu}{\sigma}$$

$$0.431 = \frac{y - 1.5}{0.18}$$

$$y = 1.58$$

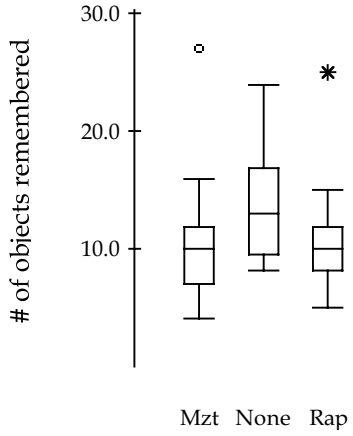
According to the Normal model, the slowest 1/3 of all drivers are expected to have reaction times of 1.58 seconds or more. (Remember that a high reaction time is a SLOW reaction time!)

26. Music and memory.

a) *Who* – 62 people. *What* – Type of music and number of objects remembered correctly. *When* – Not specified. *Where* – Not specified. *Why* – Researchers hoped to determine whether or not music affects memorization ability. *How* – Data were gathered in a completely randomized experiment.

b) Type of music (Rap, Mozart, or None) is a categorical variable. Number of items remembered is a quantitative variable.

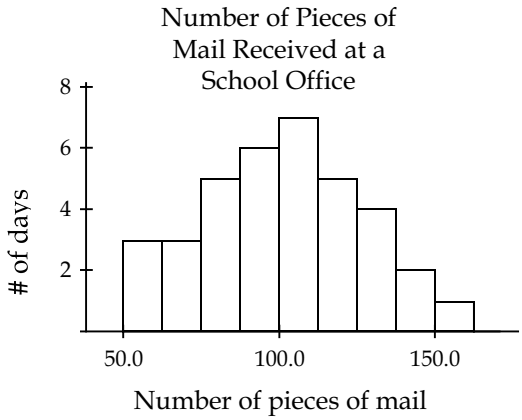
c) Accurate boxplots cannot be constructed, because we do not have all the data. By performing outlier tests, we can determine that there are no low outliers (the minimums are all within the fences), but the Rap group and the Mozart group each have at least one high outlier (the maximum in each group is above the fence). Some possible boxplots are at the right.



d) Mozart and Rap had very similar distributions of the number if objects remembered. The scores for None are, if anything, slightly higher than the other two groups. It is clear that groups listening to music (Rap or Mozart) did **not** score higher than those who listened to None.

27. Mail.

a) A histogram of the number of pieces of mail received at a school office is at the right.



b) Since the distribution of number of pieces of mail is unimodal and symmetric, the mean and standard deviation are appropriate measures of center and spread. The mean number of pieces of mail is 100.25, and the standard deviation is 25.54 pieces.

c) The distribution of the number of pieces of mail received at the school office is unimodal and symmetric, with mean 100.25 and standard deviation 25.54. The lowest number of pieces of mail received in a day was 52 and the highest was 151.

d) 23 of the 36 days (64%) had a number of pieces of mail received within one standard deviation of the mean, or within the interval 74.71 - 125.79. This is fairly close to the 68% predicted by the Normal model. The Normal model may be useful for modeling the number of pieces of mail received by this school office.

28. Birth Order.

a) There were 223 students. Of these, 113, or 50.7%, were oldest or only children.

b) There were 43 Humanities majors. Of these, 15, or 34.9%, were oldest or only children.

c) There were 113 oldest children. Of these, 15, or 13.3%, were Humanities majors.

d) There were 223 students. Of these, 15, or 6.7%, were oldest children majoring in Humanities.

29. Herbal medicine.

a) *Who* – 100 customers. *What* – Researchers asked whether or not the customer had taken the cold remedy and had customers rate the effectiveness of the remedy on a scale from 1 to 10. *When* – Not specified. *Where* – Store where natural health products are sold. *Why* – The researchers were from the Herbal Medicine Council, which sounds suspiciously like a group that might be promoting the use of herbal remedies. *How* – Researchers conducted personal interviews with 100 customers. No mention was made of any type of random selection.

b) "Have you taken the cold remedy?" is a categorical variable. Effectiveness on a scale of 1 to 10 is a categorical variable, as well, with respondents rating the remedy by placing it into one of 10 categories.

c) Very little confidence can be placed in the Council's conclusions. Respondents were people who already shopped in a store that sold natural remedies. They may be pre-disposed to thinking that the remedy was effective. Furthermore, no attempt was made to randomly select respondents in a representative manner. Finally, the Herbal Medicine Council has an interest in the success of the remedy.

30. Birth order revisited.

a) Overall, 25.6% of the students were Math/Science majors, 41.7% were Agriculture majors, 19.3% were Humanities majors, and 13.5% had other majors.

b) Of the oldest children, 30.1% of the students were Math/Science majors, 46.0% were Agriculture majors, 13.3% were Humanities majors, and 10.6% had other majors.

c) Of the second born children, 20.2% of the students were Math/Science majors, 39.1% were Agriculture majors, 24.7% were Humanities majors, and 15.9% had other majors.

d) No, college major does not appear to be independent of birth order. Oldest children are more likely than second born children to major in Math/Science (30.1% to 20.1%), while second born children are more likely than oldest children to major in Humanities (24.7% to 13.3%).

31. Engines.

a) The count of cars is 38.

b) The mean displacement is higher than the median displacement, indicating a distribution of displacements that is skewed to the right. There are likely to be several very large engines in a group that consists of mainly smaller engines.

c) Since the distribution is skewed, the median and IQR are useful measures of center and spread. The median displacement is 148.5 cubic inches and the IQR is 126 cubic inches.

d) Your neighbor's car has an engine that is bigger than the median engine, but 227 cubic inches is smaller than the third quartile of 231, meaning that at least 25% of cars have a bigger engine than your neighbor's car. Don't be impressed!

e) Using the Outlier Rule (more than 1.5 IQRs beyond the quartiles) to find the fences:

Upper Fence: Q3 + 1.5(IQR) = 231 + 1.5(126) = 420 cubic inches.

Lower Fence: Q1 – 1.5(IQR) = 105 – 1.5(126) = - 84 cubic inches.

Since there are certainly no engines with negative displacements, there are no low outliers. Q1 + Range = 105 + 275 = 380 cubic inches. This means that the maximum must be less than 380 cubic inches. Therefore, there are no high outliers (engines over 420 cubic inches).

f) It is not reasonable to expect 68% of the car engines to measure within one standard deviation of the mean. The distribution engine displacements is skewed to the right, so the Normal model is not appropriate.

g) Multiplying each of the engine displacements by 16.4 to convert cubic inches to cubic centimeters would affect measures of position and spread. All of the summary statistics (except the count!) could be converted to cubic centimeters by multiplying each by 16.4.

32. Engines, again.

a) The distribution of horsepower is roughly uniform, with a bit of skew to the right, as the number of cars begins to taper off after about 125 horsepower. The center of the distribution is about 100 horsepower. The lowest horsepower is around 60 and the highest is around 160.

b) The interquartile range is Q3 – Q1 = 125 – 78 = 47 horsepower.

c) Using the Outlier Rule (more than 1.5 IQRs beyond the quartiles) to find the fences:

Upper Fence: Q3 + 1.5(IQR) = 125 + 1.5(47) = 195.5 horsepower

Lower Fence: Q1 – 1.5(IQR) = 78 – 1.5(47) = 7.5 horsepower

From the histogram, we can see that there are no cars with horsepower ratings anywhere near these fences, so there are no outliers.

d) The distribution of horsepower is uniform, not unimodal, and not very symmetric, so the Normal model is probably not a very good model of the distribution of horsepower.

e) Within one standard deviation of the mean is roughly the interval 75 – 125 horsepower. By dividing the bars of the histogram up into boxes representing one car, and taking half of the boxes in the bars representing 70-79 and 120-129, I counted 22 (of the 38) cars in the interval. Approximately 58% of the cars are within one standard deviation of the mean.

f) Adding 10 horsepower to each car would increase the measures of position by 10 horsepower and leave the measures of spread unchanged. Mean, median, 25th percentile and 75th percentile would each increase by 10. The standard deviation, interquartile range, and range would remain the same.

33. Age and party 2007.

a) 1101 of 4002, or approximately 27.5%, of all voters surveyed were Republicans.

b) This was a representative telephone survey conducted by Gallup, a reputable polling firm. It is likely to be a reasonable estimate of the percentage of all voters who are Republicans.

c) 1001 + 1004 = 2005 of 4002, or approximately 50.1%, of all voters surveyed were under 30 or over 65 years old.

d) 409 of 4002, or approximately 10.2%, of all voters surveyed were Independents under the age of 30.

e) 409 of the 1497 Independents surveyed, or approximately 27.3%, were under the age of 30.

f) 409 of the 1001 respondents under 30, or approximately 40.9%, were Independents.

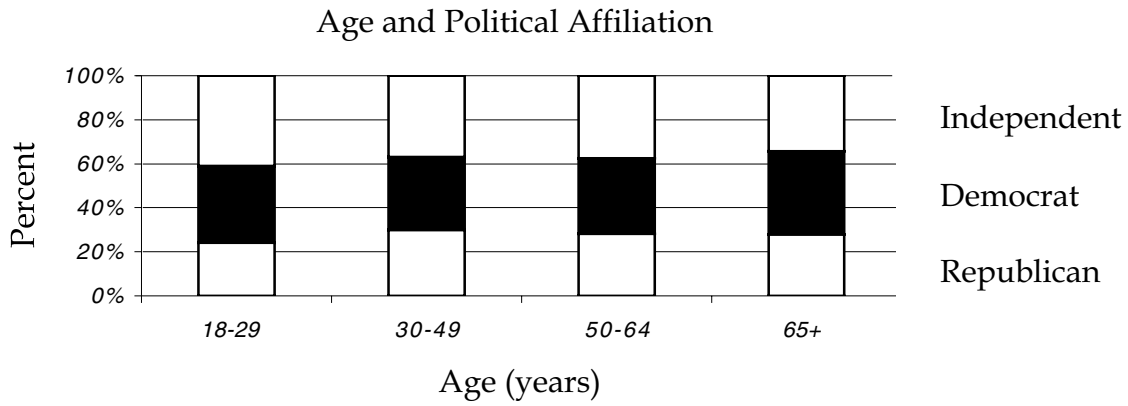
34. Pay.

The distribution of hourly wages for Chief Executives has a mean less than the median, indicating a distribution that is skewed to the left. Many Chief Executives are likely to have high hourly wages, and a few have low hourly wages, pulling the mean down. The distribution of hourly wages for General and Operations Managers has a mean higher than the median, indicating a distribution that is skewed to the right. Many General and Operations Managers have comparatively low hourly wages, and a few have high hourly wages, pulling the mean up.

35. Age and party II.

a) The marginal distribution of party affiliation is:
Republican – 27.5% Democrat – 35.1% Independent – 37.4%
(As counts: Republican – 1101 Democrat – 1404 Independent – 1497)

b)



c) Political affiliation appears to be largely unrelated to age. According to The Gallup Poll, the percentages of Independents, Democrats and Republicans within four age categories are roughly the same, with approximately 35-40% Independent, 33-38% Democrat, and 25-30% Republican. However, there is some evidence that younger voters are more likely to be Independent than older voters.

d) The percentages of Independents, Democrats, and Republicans are roughly the same within each age category. Age and political affiliation appear to be independent. At the very least, there is no evidence of a strong association between the two.

36. Bike safety 2003.

a) *Who* – Years from 1994 to 2003. *What* – Number of bicycle fatalities reported. *When* – 1994 to 2003. *Where* – United States. *Why* – The information was collected for a report by the Bicycle Helmet Safety Institute. *How* – Although not specifically stated, the information was probably collected from government agency or hospital records.

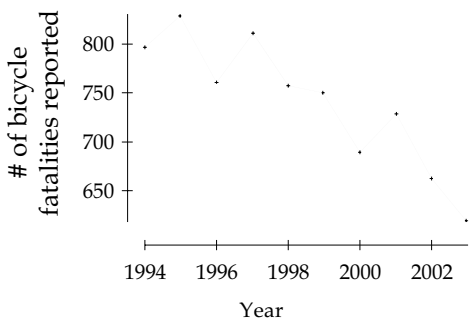
b)

United States Yearly
Bicycle Fatalities (1994 - 2003)

6	1
6	68
7	2
7	5569
8	12

6|1 = 610 - 619
Bicycle Fatalities

c)



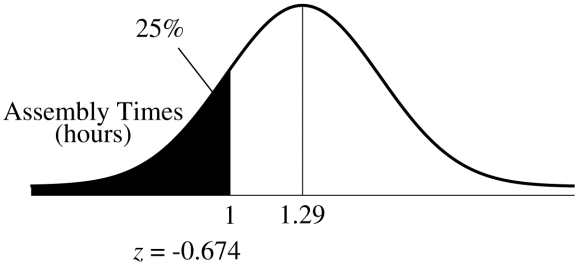
d) The stem-and-leaf display of the number of yearly bicycle fatalities reported in the United States shows that distribution is skewed to the left. It also provides some idea about the center and spread of the annual fatalities. This is not visible on the timeplot.

e) The timeplot of the number of yearly bicycle fatalities reported in the United Stats shows that the number of fatalities per year has declined over time.

f) In the years 1994 – 2003, the number of reported bicycle fatalities per year has declined fairly steadily, from approximately 800 fatalities in 1994 to approximately 620 fatalities in 2003.

37. Some assembly required.

a)



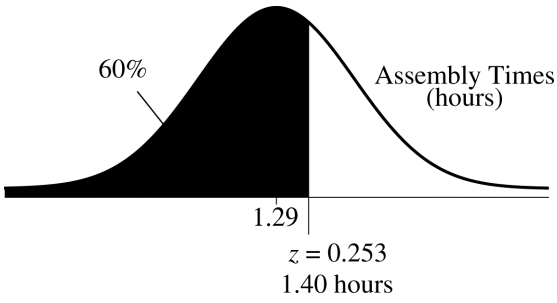
$$z = \frac{y - \mu}{\sigma}$$

$$-0.674 = \frac{1 - 1.29}{\sigma}$$

$$\sigma = 0.43$$

According to the Normal model, the standard deviation is 0.43 hours.

b)



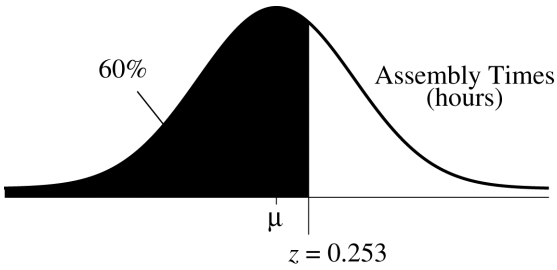
$$z = \frac{y - \mu}{\sigma}$$

$$0.253 = \frac{y - 1.29}{0.43}$$

$$y = 1.40$$

According to the Normal model, the company would need to claim that the desk takes "less than 1.40 hours to assemble", not the catchiest of slogans!

c)



$$z = \frac{y - \mu}{\sigma}$$

$$0.253 = \frac{1 - \mu}{0.43}$$

$$\mu = 0.89$$

According to the Normal model, the company would have to lower the mean assembly time to 0.89 hour (53.4 minutes).

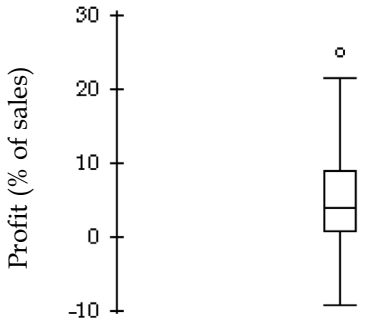
d) The new instructions and part-labeling may have helped lower the mean, but it also may have changed the standard deviation, making the assembly times more consistent as well as lower.

38. Profits.

a) The 5-number summary of the profits as a percent of sales of 29 of the *Forbes* 500 largest US corporations is: -9, 1, 4, 9.5, 25

(If you got –9, 1, 4, 9, 25, don't worry. Some statisticians figure quartiles of small sets differently than others. No one seems to care much which you use, since quartiles are much more useful in large data sets, anyway, where this doesn't matter.)

b) The boxplot of the distribution of the profits as a percent of sales of 29 of the *Forbes* 500 largest US corporations is at the right.



c) The mean profit is 4.72%, and the standard deviation of the distribution of profits is 7.55%.

d) The distribution of profits is unimodal and symmetric, centered around 4% of sales. The middle 50% of companies report profit between 1% and 9.5%. There are two companies with unusually high profits, 22% and 25%, although only 25% is technically an outlier.