Chapter 5 – Describing Distributions Numerically

1. **In the news.** Answers will vary.
2. **In the news.** Answers will vary.
3. **Time on the Internet.** Answers will vary.
4. **Groups on the Internet.** Answers will vary.
5. **Pizza prices.**

   a) Pizza 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 on average than in Dallas and Denver, but the difference is small.

   b) There are low outliers in the distribution of pizza prices in Baltimore and Chicago. There is one high outlier in the distribution of pizza prices in Dallas. These outliers do not affect the overall conclusions reached in the previous part.

6. **Costs.**

   a) Coffee is the most expensive commodity, on average.

   b) Newspapers are generally more expensive than a ride on public transportation, but there are cities in which a ride on public transportation is more expensive than a newspaper in other cities.

   c) Each distribution has a high outlier, but the presence of the outlier doesn't affect any of the previous conclusions.

7. **Still rockin'.**

   a) The histogram and boxplot of the distribution of "crowd crush" victims' ages both show that a typical crowd crush victim was approximately 18 - 20 years of age, that the range of ages is 36 years, that there are two outliers, one victim at age 36 - 38 and another victim at age 46 – 48.

   b) This histogram shows that there may have been two modes in the distribution of ages of "crowd crush" victims, one at 18 - 20 years of age and another at 22 – 24 years of age. Boxplots, in general, can show symmetry and skewness, but not features of shape like bimodality or uniformity.

   c) Median is the better measure of center, since the distribution of ages has outliers. Median is more resistant to outliers than the mean.

   d) IQR is a better measure of spread, since the distribution of ages has outliers. IQR is more resistant to outliers than the standard deviation.

8. **Slalom times.**

   a) The histogram and boxplot of the distribution of slalom times both show that a typical slalom time of around 92 seconds, that the range of slalom times is a little more than 20 seconds. Both distributions also show that the distribution of slalom times is skewed to the high end.

b) The histogram shows that the distribution of slalom times is bimodal. In addition to a mode around 92 seconds, there was a small group of skiers with scores around 105 seconds. Boxplots, in general, can show symmetry and skewness, but not features of shape like bimodality or uniformity. However, the boxplot shows two possible outlying times that the histogram doesn't highlight.

c) Since the distribution of slalom times is skewed, and contains possible outliers, the median is the better summary of center.

d) In the presence of skewness and possible outliers, we'd prefer the IQR to the standard deviation as a measure of spread.

9. Cereals.

a) The maximum sugar content is approximately 60% and the minimum sugar content is approximately 1%, so the range of sugar contents is about 60 – 1 = 59%.

b) The distribution of sugar content of cereals is bimodal, with modes centered around 5% and 45% sugar by weight.

c) Some cereals are healthy, low-sugar brands, and others are very sugary.

d) Yes. The minimum sugar content in the children's cereals is about 35% and the maximum sugar content of adult cereals is only 34%.

e) The range of sugar contents is about the same for the two types of cereals, approximately 28%, but the IQR is larger for the adult cereals. This is an indication of more variability in the sugar content of the middle 50% of adult cereals.

10. Tendon transfers.

a) The distribution of pushing strength scores is unimodal and symmetric.

b) The maximum pushing strength score is 4 and the minimum is 1, so the range is a score of 4 – 1 = 3.

c) The histogram does not show that the results of the two procedures typically resulted in different strengths.

d) The distribution of biceps transfer strength scores had a higher median than the distribution of deltoid transfer strength scores.

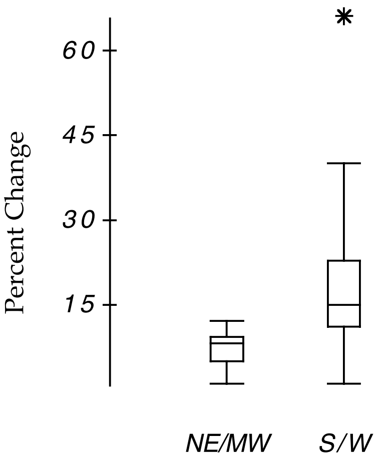
e) The biceps transfer was not always the best. The highest strength score in the deltoid transfer was higher than the lowest 25% of biceps transfer strength scores.

f) The deltoid transfer produced more consistent strength scores. The IQR is much smaller for this group, even though the ranges are approximately the same.

11. Population growth.

a) Comparative boxplots are at the right.

b) The distribution of population growth in NE/MW states is unimodal, symmetric and tightly clustered around 5% growth. The distribution of population growth in S/W states is much more spread out, with most states having population growth between 5% and 30%. A typical state had about 15% growth. There were two outliers, Arizona and Nebraska, with 40% and 66% growth, respectively. Generally, the growth rates in the S/W states were higher and more variable than the rates in the NE/MW states.


Percent Change in Population Northeastern/Midwestern vs. Southern/Western States

12. Camp sites.

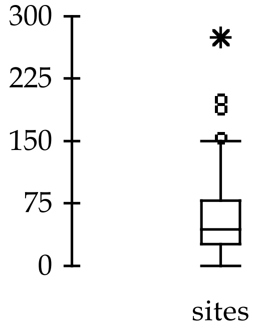
a) The distribution of the number of campsites in public parks in Vermont is skewed to the right, so median and IQR are appropriate measures of center and spread.

b) IQR = Q3 – Q1 = 78 – 28 = 50.
Using the Outlier Rule (1.5 IQRs beyond quartiles):

Upper Fence:
$$\begin{aligned} Q3+1.5(IQR) &= 78+1.5(50) \\ &= 78+75 \\ &= 153 \end{aligned}$$

Lower Fence: Well below 0 campsites.

There are 3 parks with greater than 180 campsites. These are definitely outliers. There are 2 parks with between 150 and 160 campsites each. These may be outliers as well.



c) A boxplot of the distribution of number of campsites is at the right.

d) The distribution of the number of campsites at public parks in Vermont is unimodal and skewed to the right. The center of the distribution is approximately 44 campsites. The distribution of campsites is quite spread out, with several high outliers. These parks have in excess of 150 campsites each.

13. Hospital stays.

a) The histograms of male and female hospital stay durations would be easier to compare if the were constructed with the same scale, perhaps from 0 to 20 days.

b) The distribution of hospital stays for men is skewed to the right, with many men having very short stays of about 1 or 2 days. The distribution tapers off to a maximum stay of approximately 25 days. The distribution of hospital stays for women is skewed to the right, with a mode at approximately 5 days, and tapering off to a maximum stay of approximately 22 days. Typically, hospital stays for women are longer than those for men.

c) The peak in the distribution of women's hospital stays can be explained by childbirth. This time in the hospital increases the length of a typical stay for women, and not for men.

14. Deaths 2003.

a) The distributions of ages at death of black Americans and white Americans are both skewed unimodal and skewed to the left, towards the lower ages of death.

b) The distributions of death ages differ mainly in their spreads and centers. The age at death for black Americans is more variable than the age at death for white Americans. A greater proportion of black Americans die between the ages of 25 and 74 than do white Americans, while a greater proportion of white Americans dies at ages older than 74 than do black Americans.

c) The interval widths are not consistent. Most of the bars are 10 years wide, but the first bar is only 5 years wide, while the last bar has width "85 and older".

15. Women's basketball.

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 about as often.

16. Gas prices.

a) Gas prices have been increasing on average over the three year period, and the spread has been increasing as well. The distribution of prices in 2002 was skewed to the left with several low outliers. Since then, the distribution has been increasingly skewed to the right. There is a high outlier in 2004, although it appears to be pretty close to the upper fence.

b) The distribution of gas prices in 2004 shows the greatest range and the biggest IQR, so the prices varied a great deal.

17. Marriage age.

The distribution of marriage age of U.S. men is skewed right, with a typical man (as measured by the median) first marrying at around 24 years old. The middle 50% of male marriage ages is between about 23 and 26 years. For U.S. women, the distribution of marriage age is also skewed right, with median of around 21 years. The middle 50% of female marriage age is between about 20 and 23 years. When comparing the two distributions, the most striking feature is that the distributions are nearly identical in spread, but have different centers. Females typically seem to marry earlier than males. In fact, between 50% and 75% of the women marry at a younger age than *any* man.

18. Fuel economy.

Cars with 4 cylinders generally get better gas mileage than cars with 6 cylinders, which generally get better gas mileage than cars with 8 cylinders. Additionally, the greater the number of cylinders, the more consistent the mileage becomes. 4 cylinder cars typically get between 27 – 33 mpg, 6 cylinder cars typically get between 18 – 22 mpg, and 8 cylinder cars typically get between 16 – 19 mpg. Cars with 5 cylinders typically got about 20 mpg.

19. Fuel economy II.

(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 get 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 from vans is about 10 mpg, while for SUVS it is nearly 30 mpg.

20. Ozone.

a) April had the highest recorded ozone level, approximately 440.

b) February had the largest IQR of ozone level, approximately 50.

c) August had the smallest range of ozone levels, approximately 50.

d) January had a slightly lower median ozone level than June, 340 and 350, respectively, but June's ozone levels were much more consistent.

e) Generally, ozone levels rose through the winter and were highest in the spring, then fell through the summer and were lowest in the fall. Additionally, ozone levels were very consistent in the summer, became more variable in the fall, were most variable in the winter, and became more consistent through the spring.

21. Test scores.

Class A is Class 1. The median is 60, but has less spread than Class B, which is Class 2. Class C is Class 3, since it's median is higher, which corresponds to the skew to the left.

22. Eye and hair color.

The graph is not appropriate. Boxplots are for quantitative data, and these are categorical data, although coded as numbers. The numbers used for hair color and eye color are arbitrary, so the boxplot and any accompanying statistics for eye color make no sense.

23. Graduation?

a) The distribution of the percent of incoming college freshman who graduate on time is roughly symmetric. The mean and the median are reasonably close to one another and the quartiles are approximately the same distance from the median.

b) Upper Fence:

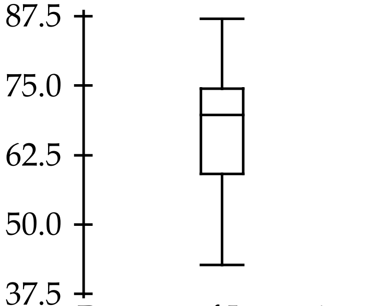
$$\begin{aligned} Q3 + 1.5(IQR) &= 74.75 + 1.5(74.75 - 59.15) \\ &= 74.75 + 23.4 \\ &= 98.15 \end{aligned}$$

Lower Fence:

$$\begin{aligned} Q1 - 1.5(IQR) &= 59.15 - 1.5(74.75 - 59.15) \\ &= 59.15 - 23.4 \\ &= 35.75 \end{aligned}$$

Since the maximum value of the distribution of the percent of incoming freshmen who graduate on time is 87.4% and the upper fence is 98.15%, there are no high outliers. Likewise, since the minimum is 43.2% and the lower fence is 35.75%, there are no low outliers. Since the minimum and maximum percentages are within the fences, all percentages must be within the fences.

c) A boxplot of the distribution of the percent of incoming freshmen who graduate on time is at the right.



Percent of Incoming Freshman Graduating in 4 Years

d) The distribution of the percent of incoming freshmen who graduate on time is roughly symmetric, with mean of approximately 68% of freshmen graduating on time. Universities surveyed had between 43.2% and 87.4% of students graduating on time, with the middle 50% of universities reporting between 59.15% and 74.75% graduating on time.

24. Vineyards.

a) The distribution of size of Finger Lakes vineyards is skewed heavily to the right. The mean size is a great deal higher than the median size.

b) Upper Fence:

$$\begin{aligned} Q3 + 1.5(IQR) &= 55 + 1.5(55 - 18.5) \\ &= 55 + 54.75 \\ &= 109.75 \end{aligned}$$

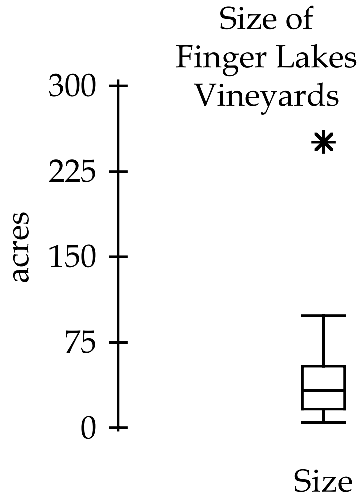
Lower Fence:

$$\begin{aligned} Q1 - 1.5(IQR) &= 18.5 - 1.5(55 - 18.5) \\ &= 18.5 - 54.75 \\ &= -36.25 \end{aligned}$$

The maximum of 250 acres is well above the upper fence of 109.75 acres. Therefore, there is at least one high outlier, 250 acres. Since the lower fence is negative, there are no low outliers, since it is certainly impossible to have a vineyard with negative size.

c) The boxplot of the distribution of sizes of Finger Lakes vineyards is at the right. There may be additional outliers, but we are sure that there is at least one, the maximium.

d) The distribution of sizes of Finger Lakes vineyards is skewed to the right. Many vineyards have moderate sizes, with the middle 50% of vineyards consisting of 18.5 to 55 acres. The smallest vineyard is 6 acres. At least one vineyard is comparatively bigger, at 250 acres.

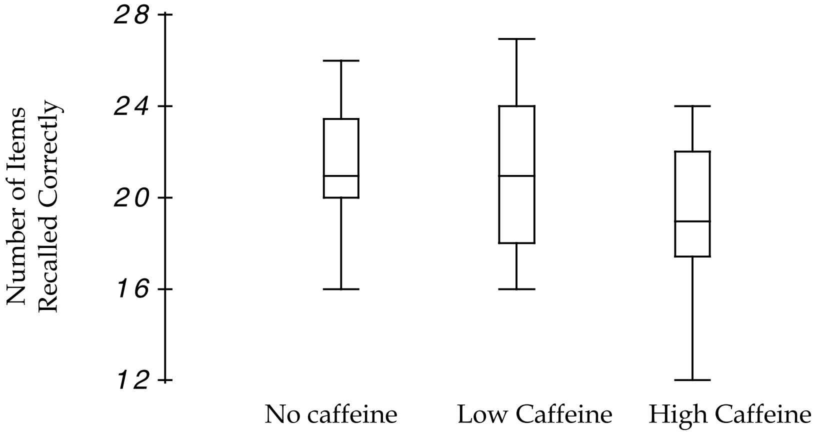


25. Caffeine.

a) *Who* – 45 volunteeers. *What* – Level of caffeine consumption and memory test score. *When* – Not specified. *Where* – Not specified. *Why* – The student researchers want to see the possible effects of caffeine on memory. *How* – It appears that the researchers imposed the treatment of level of caffeine consumption in an experiment. However, this point is not clear. Perhaps they allowed the subjects to choose their own level of caffeine.

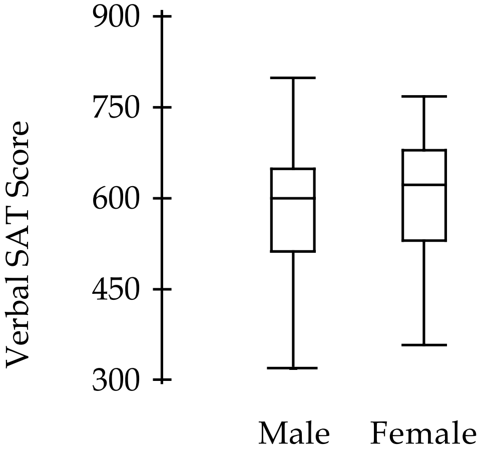
b) *Variables* – Caffeine level is a categorical variable with three levels: no caffeine, low caffeine, and high caffeine. Test score is a quantitative variable, measured in number of items recalled correctly.

c)



d) The groups consuming no caffeine and low caffeine had comparable memory test scores. A typical score from these groups was around 21. However, the scores of the group consuming no caffeine were more consistent, with a smaller range and smaller interquartile range than the scores of the group consuming low caffeine. The group consuming high caffeine had lower memory scores in general, with a median score of about 19. No one in the high caffeine group scored above 24, but 25% of each of the other groups scored above 24.

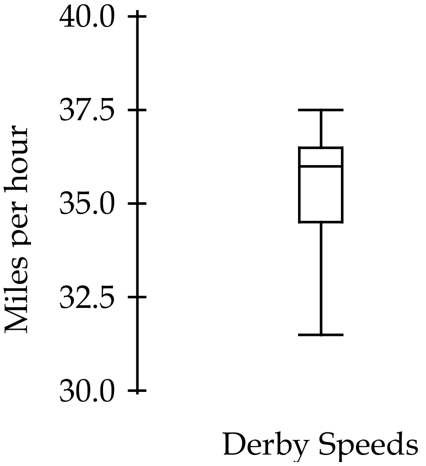
26. SAT scores.



a) Parallel boxplots comparing the scores of boys and girls SAT scores are at the right.

b) Females in this graduating class scored slightly higher on the Verbal SAT, with a median of 625, compared to the median of 600 for the males. Additionally, the females had higher first and third quartiles. The IQR of the males' scores was slightly smaller, than the IQR for the females' scores, indicating a bit more consistency in male scores. However, the overall spread of male scores was greater than that of female scores, with males having both the minimum and maximum score. Both distributions of scores were slightly skewed to the left.

27. Derby speeds 2007.



a) The median speed is the speed at which 50% of the winning horses ran slower. Find 50% on the left, move straight over to the graph and down to a speed of about 36 mph.

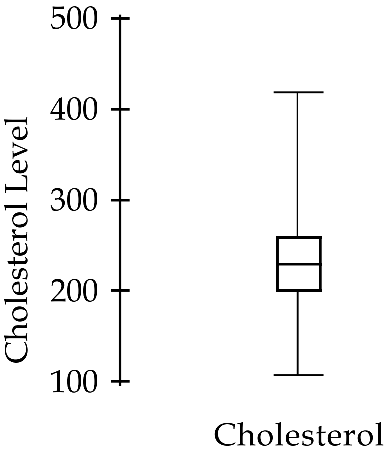
b) Quartile 1 is at 25% on the left, and Quartile 3 is at 75% on the left. Matching these to the ogive, Q1 = 34.5 mph and Q3 = 36.5 mph.

c) Range = Max – Min = 38 – 31 = 7 mph
IQR = Q3 – Q1 = 36.5 – 34.5 = 2 mph

d) An approximate boxplot of winning Kentucky Derby Speeds is at the right.

e) The distribution of winning speeds in the Kentucky Derby is skewed to the left. The lowest winning speed is just under 31 mph and the fastest speed is about 37.5 mph. The median speed is approximately 36 mph, and 75% of winning speeds are above 34.5 mph. Only a few percent of winners have had speeds below 33 mph.

28. Cholesterol.



A boxplot for the distribution of cholesterol levels of 1400 men is at the right. The five number summary is estimated from the ogive to be: 100, 200, 230, 260, 425.

The distribution of cholesterol levels is skewed to the right, and tightly clustered around the median. The median cholesterol level is approximately 230, with the middle 50% of cholesterol levels between 200 and 260. A very small percentage of men had cholesterol below 150 or above 325.

29. Reading scores.

a) The highest score for boys was 6, which is higher than the highest score for girls, 5.9.

b) The range of scores for boys is greater than the range of scores for girls.
Range = Max – Min   Range(Boys) = 4   Range(Girls) = 3.1

c) The girls had the greater IQR.
IQR = Q3 – Q1   IQR(Boys) = 4.9 – 3.9 = 1   IQR(Girls) = 5.2 – 3.8 = 1.4

d) The distribution of boys' scores is more skewed. The quartiles are not the same distance from the median. In the distribution of girls' scores, Q1 is 0.7 units below the median, while Q3 is 0.7 units above the median.

e) Overall, the girls did better on the reading test. The median, 4.5, was higher than the median for the boys, 4.3. Additionally, the upper quartile score was higher for girls than boys, 5.2 compared to 4.9. The girls' lower quartile score was slightly lower than the boys' lower quartile score, 3.8 compared to 3.9.

f) The overall mean is calculated by weighting each mean by the number of students.

$$\frac{14(4.2)+11(4.6)}{25} = 4.38$$

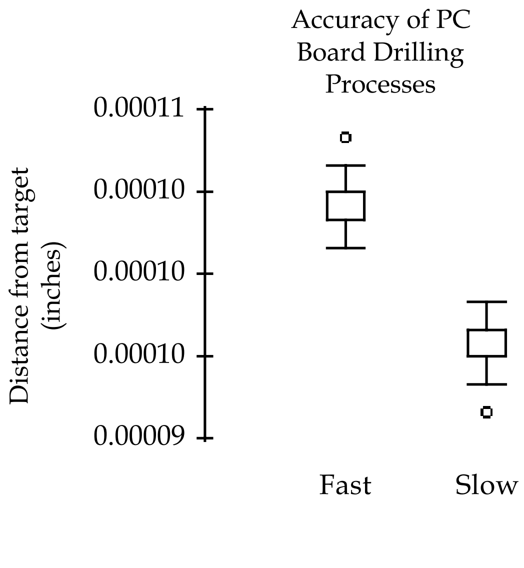
30. Rainmakers?

a) Median and IQR, as well as the quartiles are the appropriate summary statistics, since the distribution of amount of rain produced is either skewed to the right or has high outliers. Indication of skewness and outliers can be seen in the comparison of median and mean. The mean amount of rain produced is significantly higher than the median for both seeded and unseeded clouds. Skewness or outliers pulled up the sensitive mean.

b) There is evidence that that the seeded clouds produced more rain. The median and both quartiles are higher than the corresponding statistics for unseeded clouds. In fact, the median amount of rainfall for seeded clouds is 221.60 acre-feet, about 5 times the median amount for unseeded clouds.

31. Industrial experiment.

First of all, there is an extreme outlier in the distribution of distances for the slow speed drilling. One hole was drilled almost an inch away from the center of the target! If that distance is correct, the engineers at the computer production plant should investigate the slow speed drilling process closely. It may be plagued by extreme, intermittent inaccuracy. The outlier in the slow speed drilling process is so extreme that no graphical display can display the distribution in a meaningful way while including that outlier. That distance should be removed before looking at a plot of the drilling distances.



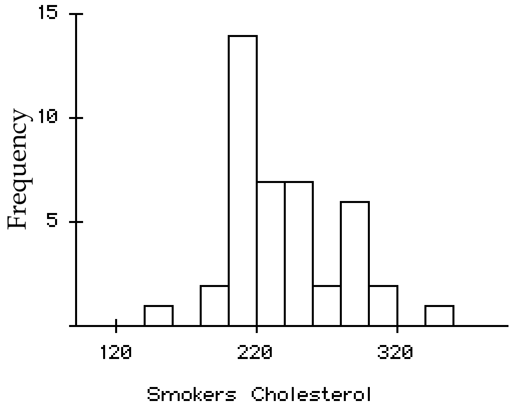
With the outlier set aside, we can determine that the slow drilling process is more accurate. The greatest distance from the target for the slow drilling process, 0.000098 inches, is still more accurate than the smallest distance for the fast drilling process, 0.000100 inches.

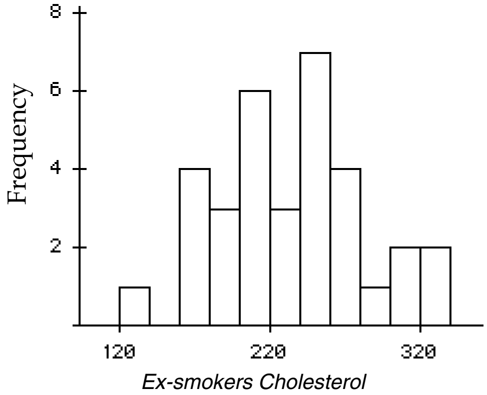
32. Cholesterol.

The distribution of cholesterol levels for smokers is unimodal and skewed slightly to the right, with a mode around 210. Cholesterol levels vary from approximately 140 to 350, but are generally clustered between 200 and 300. There is one low cholesterol level and one high cholesterol level, but these don't depart from the overall pattern.

The distribution of cholesterol levels for ex-smokers is unimodal and roughly symmetric, with a center around 240. Cholesterol levels vary from approximately 120 to 340, and seem spread out. There is one value, but not unusually low.

In general, the cholesterol levels of smokers seem to be slightly lower than the cholesterol levels of ex-smokers. Additionally, the cholesterol levels of smokers appear more consistent than cholesterol levels of ex-smokers.





33. MPG.

a) A back-to-back stemplot of these data is shown at the right. A plot with with tens and units digits for stems and tenths for leaves would have been quite long, but still useable. A plot with tens as stems and rounded units as leaves would have been too compact. This plot has tens as the stems, but the stems are split 5 ways. The uppermost 2 stem displays 29 and 28, the next 2 stem displays 27 and 26, and so on. The key indicates the rounding used, as well as the accuracy of the original data. In this case, the mileages were given to the nearest tenth.

US Cars		Other Cars
	3	7
44	3	45
	3	222
01	3	01
89	2	8
777	2	7
	2	
2	2	222
11	2	0
888999	1	
6777	1	67

KEY:
2 | 5 = 24.5 – 25.4 mpg

b) In general, the Other cars got better gas mileage than the US Cars, although both distributions were highly variable. The distribution of US cars was bimodal and skewed to the right, with many cars getting mileages in the high teens and low twenties, and another group of cars whose mileages were in the high twenties and low thirties. Two high outliers had mileages of 34 miles per gallon. The distribution of Other cars, in contrast, was bimodal and skewed to the left. Most cars had mileages in the high twenties and thirties, with a small group of cars whose mileages were in the low twenties. Two low outliers had mileages of 16 and 17 miles per gallon.

34. Baseball.

a) The back-to-back stemplot shown at the right has split stems to show the distribution in a bit more detail than a stemplot with single stems.

American		National
	14	0
	13	
	13	
	12	
	12	
	11	6
1	11	
88	10	
0133	10	234
5	9	5555
02234	9	1
	8	8
3	8	1234
	7	9

KEY:
10 | 3 = 10.3
runs per game

b) The distribution of number of runs per game in American League stadiums is unimodal and slightly skewed to the right, clustered in the interval of 9 to 10 runs per game. In the National League, the number of runs scored per game is distributed symmetrically and is possibly bimodal, with clusters in the low 8s, high 9s, and low 10s a. There are two high outliers of 11.6 and 14 runs per game. The number of runs scored per game is generally higher and more consistent in the American League.

c) The 14 runs per game scored at Coors Field is an outlier in the National League data for the first half of the 2001 season. There appear to be more runs per game scored there than in other Major League Stadiums.

35. Fruit Flies.

a) The most fruit flies died around day 16. Over 60,000 fruit flies died that day.

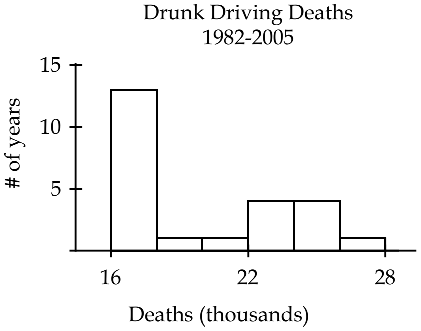
b) The largest proportion of fruit flies died around day 65. About 20% of that previously surviving population died that day.

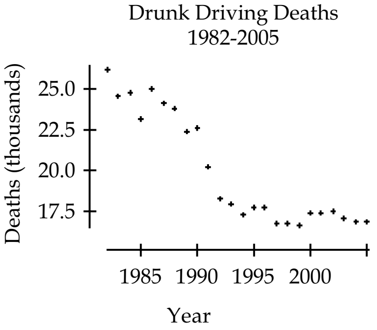
c) At around day 50, the number of fruit flies wasn't changing by very much day to day.

36. Drunk driving 2005.

a) The histogram (near right) shows the distribution of drunk driving deaths.

b) The timeplot (far right) shows the change in drunk driving deaths over time.





c) The distribution of the number of drunk driving deaths is bimodal, with a cluster between 22 and 25 thousand deaths and another cluster between 16 and 17 thousand deaths. The timeplot shows that this corresponds to a rapid decrease in the drunk driving deaths in the early nineties. The number of deaths was high, then decreased dramatically. Since about 1995, the number of drunk driving deaths has leveled off.

37. Assets.

a) The distribution of assets of 79 companies chosen from the *Forbes* list of the nation's top corporations is skewed so heavily to the right that the vast majority of companies have assets represented in the first bar of the histogram, 0 to 10 billion dollars. This makes meaningful discussion of center and spread impossible.

b) Re-expressing these data by, for example, logs or square roots might help make the distribution more nearly symmetric. Then a meaningful discussion of center might be possible.

38. Music library.

a) The distribution of the number of songs students had in their digital libraries is extremely skewed to the right. That makes it difficult to determine a center. The typical number of songs in a library is probably in the first bar of the histogram.

b) Re-expressing these data by, for example, logs or square roots might help make the distribution more nearly symmetric. Then a meaningful discussion of center might be possible.

39. Assets again.

a) The distribution of logarithm of assets is preferable, because it is roughly unimodal and symmetric. The distribution of the square root of assets is still skewed right, with outliers.

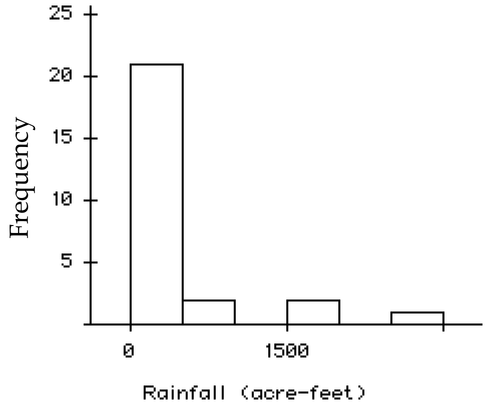
b) If $\sqrt{Assets} = 50$, then the companies assets are approximately $50^2 = 2500$ million dollars.

c) If $\log(Assets) = 3$, then the companies assets are approximately $10^3 = 1000$ million dollars.

40. Rainmakers.

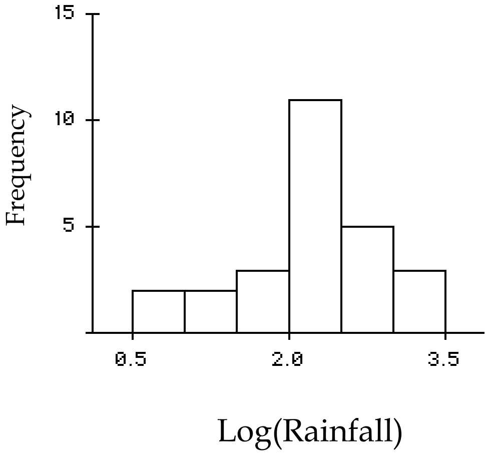
a) Since one acre-foot is about 320,000 gallons, these numbers are more manageable than gallons.

b) The distribution of rainfall from 26 clouds seeded with silver iodide is skewed heavily to the right, with the vast majority of clouds producing less than 500 acre-feet of rain. Several clouds produced more, with a maximum of 2745 acre-feet.



c) The distribution of log (base 10) of rainfall is much more symmetric than the distribution of rainfall. We can see that the center of the distribution is around log 2 – log 2.5 acre-feet.

d) Since the reexpressed scale is measured in log (acre-feet), we need to raise 10 to the power of the number on our scale to convert back to acre feet. For example, if a cloud in the new scale has a log (rainfall) of 2.3, we convert back to rainfall as follows:



$$\begin{aligned}\log(rainfall) &= 2.3\\ rainfall &= 10^{2.3}\\ rainfall &= 199.5\end{aligned}$$

The cloud produced 199.5 acre-feet of rain.

41. Stereograms.

a) The two variables discussed in the description are fusion time and treatment group.

b) Fusion time is a quantitative variable, measured in seconds. Treatment group is a categorical variable, with subjects either receiving verbal clues only, or visual and verbal clues.

c) Generally, the Visual/Verbal group had shorter fusion times than the No/Verbal group. The median for the Visual/Verbal group was approximately the same as the lower quartile for the No/Verbal group. The No/Verbal Group also had an extreme outlier, with at least one subject whose fusion time was approximately 50 seconds. There is evidence that visual information may reduce fusion time.

42. Stereograms, revisited.

The re-expression using logarithms has a distribution that is more symmetric than the original distribution of fusion times, and the re-expression has no outliers. This symmetry makes it easier to compare the two groups.