Complete Solutions Manual

Mind on Statistics

FIFTH EDITION

Jessica M. Utts
University of California, Irvine
Irvine, CA

Robert F. Heckard
Pennsylvania State University
State College, PA

CENGAGE Learning®

Australia • Brazil • Mexico • Singapore • United Kingdom • United States

ISBN-13: 978-130510245-3
ISBN-10: 1-305-10245-2

Cengage Learning
200 First Stamford Place, 4th Floor
Stamford, CT 06902
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil, and Japan. Locate your local office at: **www.cengage.com/global**.

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

To learn more about Cengage Learning Solutions, visit **www.cengage.com**.

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com**.

Printed in the United States of America
1 2 3 4 5 6 7 17 16 15 14 13

Contents

Chapter 1 3

Chapter 2 9

Chapter 3 45

Chapter 4 65

Chapter 5 87

Chapter 6 100

Chapter 7 114

Chapter 8 134

Chapter 9 153

Chapter 10 193

Chapter 11 211

Chapter 12 235

Chapter 13 265

Chapter 14 296

Chapter 15 307

Chapter 16 332

Chapter 17 345

CHAPTER 1
EXERCISE SOLUTIONS

1.1 **a.** The fastest speed was 150 miles per hour.
b. The slowest speed driven by a male was 55 miles per hour.
c. 1/4 of the females reported having driven at 95 miles per hour or faster. Notice that 95 mph is the *upper quartile* for females. By definition, about 1/4 of the values in a data set are greater than the upper quartile.
d. 1/2 of the females reported having driven 89 mph or faster. Notice that 89 mph is the *median* value.
e. 1/2 of 102 = 51 females have driven 89 mph or faster.
Note: For parts (d) and (e) the answer would have to be adjusted if there were any females who reported 89 as their value, but from the data on page 2 we can see that there were not. Because there were no "ties" with the median, we know that exactly half of the values fall above it and half fall below it.

1.2 **a.** The median height is 65 inches.
b. Range = Tallest – Shortest = 71 – 59 = 12 inches.
c. The interval from 59 to 63.5 inches contains the shortest 1/4 of the women. This interval is from the minimum to the lower quartile.
d. The interval from 63.5 to 67.5 inches contains the middle 1/2 of the women. This is an interval from the lower quartile to the upper quartile.

1.3 **a.** The observed rate of cervical cancer in Vietnamese American women is 86 per 200,000. This could also be expressed as 43 per 100,000 or 4.3 per 10,000, and so on. In decimal form, it is .00043.
b. The risk of developing cervical cancer for Vietnamese American women in the next year is 86/200000 = .00043.
c. The rate of 86 per 200,000 is based on past data and tells us the number of Vietnamese American who developed cervical cancer out of a population of 200,000. The risk utilizes the rate from the past to tell us the future likelihood of cervical cancer in other Vietnamese American women.

1.4 **a.** The base rate is about 13 in 1000, or about .013.
b. The risk for men who smoke is just over 13 times the rate for non-smokers, or about .169.

1.5 **a.** All teens in the U.S. at the time the poll was taken.
b. All teens in the U.S. who had dated at the time the poll was taken.

1.6 A population is a collection of all individuals of interest while a sample is a subset of the population of interest, for which measurements are taken in a study. In Case Study 1.6 the population of interest is probably all men, and possibly women as well. However, the sample consisted of 22,071 male physicians who volunteered for the study, so the population to which the results apply is all men similar to them.

1.7 **a.** All adults in the U.S. at the time the poll was taken.
b. $\frac{1}{\sqrt{1048}} = .031$ or 3.1%
c. 34% ± 3.1%, or 30.9% to 37.1%.

1.8 **a.** The population is probably all Canadians who were eligible to participate in the survey (which is probably all adults with telephones).
b. There were 2000 people in the sample.
c. $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{2000}} = .022$ or 2.2%.
d. In the sample, 16% viewed immigration as having a negative impact. The interval 16% ± 2.2%, or 13.8% to 18.2% is 95% certain to cover the true percent of Canadians who viewed immigration as having a negative impact (at the time of the poll).

1.9 Solve for n in the equation $\frac{1}{\sqrt{n}} = .05 = \frac{1}{20}$. Answer is $n = 400$ teenagers.

1.10 Solve for n in the equation $\frac{1}{\sqrt{n}} = .30$. "Exact" answer is 11.11, which is not a possible sample size, so round up to $n = 12$ to guarantee a margin of error that's less than 30%. With $n = 11$, margin of error = 30.1%.

1.11 **a.** This is an example of a self-selected or volunteer sample. Magazine readers voluntarily responded to the survey, and were not randomly selected.
b. These results may not represent the opinions of all readers of the magazine. The people who respond probably do so because they feel stronger about the issues (for example, violence on television or physical discipline) than the readers who do not respond. So, they may be likely to have a generally different point of view than those who do not respond.

1.12 The exercise did not specify what the survey is about, but no matter what, the survey is based on a self-selected sample, and people who feel strongly about the issues and/or who have extra time are more likely to respond. The results will not be representative of all students who use the cafeteria.

1.13 **a.** Randomized experiment (because students were randomly assigned to the two methods).
b. Observational study (because people cannot be randomly assigned to smoke or not).
c. Observational study (because people cannot be randomly assigned to be a CEO or not).

1.14 **a.** Randomized experiment, because students were randomly assigned to receive Vitamin C or placebo.
b. Observational study, because the patients are not randomly assigned to do anything. (Note that a random sample is not the same thing as random assignment.)
c. Randomized experiment, because participants were randomly assigned to meditation or low-fat diet.

1.15 Answers will vary, but one possibility is general level of activity. It is likely to differ for elderly people who attend church regularly and those who don't, and it is also likely to affect blood pressure. So it might partially explain the results of this study.

1.16 **a.** Number of courses might be a confounding variable. Students taking many courses may sleep less due to the amount of work involved and may not do as well in school due to the load.
b. Weight is not likely to be a confounding variable. Weight is probably not related either to amount of sleep or to grades.
c. Hours spent partying might be a confounding variable. Students who party a lot may sleep less, and may also get lower grades because they're not studying.

1.17 You would need to know how large the difference in weight loss was for the two groups. If the difference in weight loss is very small (but not 0) it could be statistically significant, but not have much practical importance.

1.18 Statistical significance is when there is a relationship or difference that is large enough to be unlikely to have occurred in the sample if there was no relationship or difference in the population of interest. Practical significance occurs when the relationship or difference is large enough to be important or meaningful in a "real world" sense. A result can be statistically significant, but not practically significant. This may occur in studies with very large sample sizes.

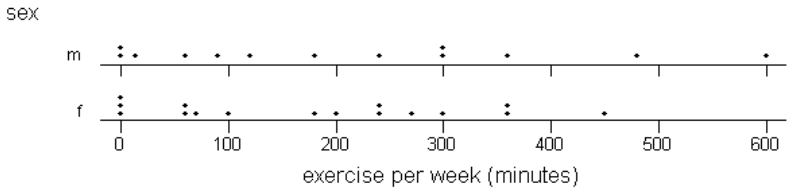
1.19 You would want to know how many different relationships were examined. If this result was the only one that was statistically significant out of many examined, it could easily be a false positive.

1.20 A false positive occurs when a relationship or difference is said to be statistically significant based on examining information from a sample, but in fact there is no relationship or difference in the population.

1.21 The placebo group estimates the baseline rate of heart attacks for men not taking aspirin. So, the estimated baseline rate of heart attacks is 189/11,034, which is about 17 heart attacks per 1,000 men or 17/1000. (See Table 1.1 for the data.)

1.22 **a.** The amount of exercise per week is similar for men and women except that there are a few high values for the men. The dotplot follows.

Figure for Exercise 1.22a



b. *Women*, median = 190 minutes. *Men*, median = 180 minutes.
To find the median, put the data in order first.
For *women*, the ordered list of data is:

0, 0, 0, 60, 60, 70, 100, **180**, **200**, 240, 240, 270, 300, 360, 360, 450

The number of women is even (16), so the median is the average of the middle two values in the ordered list. These middle two values, underlined and bold in the above list, are 180 and 200 and their average is 190.
For *men*, the ordered list of data is:

0, 0, 14, 60, 90, 120, **180**, 240, 300, 300, 360, 480, 600

The number of men is odd (13) so the median is the middle value in the ordered data. This value, underlined and bold in the list above, is 180.
c. Although the median response is different for women and men, the difference is only 10 minutes. The weekly amount of exercise is about the same for the samples of women and men.

1.23 **a.**

	Minutes of exercise per week		
Median		180	
Quartiles	37		330
Extremes	0		600

To determine the summary, first write the responses in order from smallest to largest.
The ordered list of data is:

0, 0, 14, 60, 90, 120, 180, 240, 300, 300, 360, 480, 600

Minimum = 0 min.
Maximum = 600 min.
Median = 180 min. (middle value in the ordered list)
Lower quartile = 42 min. It is the median of the values smaller than the median.
These are 0, 0, 14, 60, 90, 120.
Median of these six values is (14+70)/2 = 42.
Upper quartile = 330 min. It is the median of the values larger than the median.
Values larger than the median are 240, 300, 300, 360, 480, 600.
Median of these values is (300+360)/2 = 330.
b. Reported exercises hours per week for the men in the sample ranged from a low of 0 to a high of 600 minutes per week. The median response was 180 min (3 hours). About 1/2 of the men (the middle half) reported exercising between 37 and 330 minutes (5 and a half hours) per week. About 1/4 said they exercised less than 37 minutes per week while 1/4 said they exercised more than 330 minutes per week.

1.24 **a.**

	Minutes of exercise per week		
Median		190	
Quartiles	60		285
Extremes	0		450

To determine the summary, first write the responses in order from smallest to largest.
The ordered list of data is:

0, 0, 0, 60, 60, 70, 100, 180, 200, 240, 240, 270, 300, 360, 360, 450

Minimum = 0 min.
Maximum = 450 min.
Median = 190 min. (average of the middle two values in the ordered list, which are 180 and 200)
Lower quartile = 60 min. It is the median of the values smaller than the median.
These are 0, 0, 0, 60, 60, 70, 100, 180.
Median of these eight values is (60+60)/2 = 60.
Upper quartile = 285 min. It is the median of the values larger than the median.
Values larger than the median are 200, 240, 240, 270, 300, 360, 360, 450.
Median of these values is (270+300)/2=285.
b. Reported exercise hours for the women in the sample ranged from a low of 0 to a high of 450 minutes per week. The median response was 190 min (3 hours and 10 minutes). About 1/2 of the women (the middle half) reported exercising between 60 and 285 minutes per week. About 1/4 said they exercised less than 60 minutes per week while 1/4 said they exercised more than 285 minutes per week.

1.25 **a.** This is an observational study because vegetarians and non-vegetarians are compared and these groups occur naturally. People were not assigned to treatment groups.
b. Since this is an observational study and not a randomized experiment, we cannot conclude that a vegetarian diet causes lower death rates from heart attacks and cancer. Other variables not accounted for may be causing this reduction.
c. This answer will differ for each student. One potential confounding variable is amount of exercise. This is a confounding variable because it may be that vegetarians also exercise more on average and this led to lower death rates from heart attacks and cancer.

1.26 Base rates were not given. In this study, a base rate would be the actual rate (risk) of a particular cause of death for people who are not vegetarians.

1.27 The base rate or baseline risk is missing from the report. You need to know the base rate of cancer of the rectum for men to decide if the increased risk from drinking beer is large or small.

1.28 **a.** This was a randomized experiment because volunteers were randomly assigned to wear either a nicotine patch or a placebo patch.
b. You can conclude that use of nicotine patches leads to a higher success rate for those trying to quit smoking than use of placebo patches.
c. It was advisable to assign some of the patients to wear a placebo patch because then you can compare the success rate of those patients to the success rate of the patients wearing nicotine patches. You will also learn in a future chapter that even though they have no active ingredients, placebos can have a large psychological effect. Also, presumably people in the experiment want to quit smoking, so some will succeed regardless of treatment method.

1.29 For Caution 1: Because the difference given in the previous exercise is a large difference (46% with nicotine patch and 20% with placebo), it has practical importance as well as statistical significance. For Caution 2: Because the result is based on a randomized experiment, it is not possible that whether someone quit or not influenced the type of patch they were assigned.

1.30 In the study described in Exercise 1.28, data were gathered and analyzed in order to make a decision about the effectiveness of the nicotine patch. This information will help individuals decide whether to wear nicotine patches when trying to quit smoking. Although the observed result was based only on a sample from a larger population, the data collection and analysis methods make it reasonable to conclude that the patch is more effective than a placebo for the population represented by this sample.

1.31 Neither caution applies. The magnitude of the difference is given in Case Study 1.6, and considering the number of men between the ages of 40 and 84 in the United States population, the given difference has

practical importance. Because men were randomly assigned to take aspirin (or not) we can conclude that the correct direction of the cause and effect is that taking aspirin caused the reduction in heart attacks

1.32 **a.** The population is all University of California faculty members in 1995.

b. The margin of error is approximately $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1000}} = .032$, or roughly 3%.

c. It cannot be concluded that a majority of all University faculty favored the criteria. In the sample a slight majority (52%) was in favor, but the margin of error was about 3%. So, in the population the percent in favor could possibly be a minority (below 50%). An interval that is 95% certain to contain the percent in the population is 52% ± 3%, which is 49% to 55%. Because some values within this interval are below 50%, we are not able to rule out the possibility that the percent in favor in the population is a minority.

1.33 **a.** The margin of error is about $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{1525}} = .026$.

b $.139 \pm .026$, which is .113 to .165.
This is *sample proportion* ± *margin of error*.

1.34 For someone 18 to 29 years old, the risk of seeing a ghost is 212/1525 = .139.

1.35 **a.** This is a self-selected (or volunteer) sample.
b. Probably higher, because people who would say they have seen a ghost would be more likely to call the late-night radio talk show than others. They might even be more likely to be listening to such a show.

1.36 **a.** Self-selected sample or volunteer sample.
b. The results cannot be extended to any larger population because the sample was not selected to be representative of any population.

1.37 The term "data snooping" refers to looking at the data in a variety of ways until something interesting to report emerges.

1.38 **a.** This statement has to be based on an observational study. The researchers observed who was breast-fed and who was not. It would not be possible to randomly assign mothers to breast-feed or not. It would not be ethical to randomly assign this treatment.
b. The better headline is "Link found between breast-feeding and school performance" because it does not imply that there is a cause-and-effect conclusion, while the other headline does. Because the study was observational, a cause-and-effect conclusion cannot be made. It is likely that children who were breast-fed as infants have other (confounding) factors in their lives that differ from children who were not, and that may influence school performance. For example, perhaps they are more likely to have mothers who don't work, perhaps they are more likely to be first-born children, and so on. We can say that a link was found, but cannot say that breast-feeding *leads* to better performance.

1.39 In some situations it is not practical or even possible to conduct a randomized experiment. For example, a researcher may wish to study whether occupational exposure to asbestos affects the risk of lung disease. It would not be possible, or ethical, to assign people to occupations that involve differing amounts of exposure to asbestos.

1.40 An observational study was done instead of an experiment because the researchers could not assign individuals either to attend a religious service once a week and pray regularly or to not engage in these practices.

1.41 The answer will differ for students, but here is an example. Randomly assign volunteers to either eat lots of chocolate or not eat any chocolate for a period of time, and give them a questionnaire about depression at the beginning and the end of the time period. Then compare the change in depression scores for the two groups.

1.42 **a.** Randomized experiment. People would not take a placebo as part of an observational study.
b. Maybe not, because 20 side effects were tested so one or a few could appear to be a problem (statistically significant) just by chance even if none of them were a problem. In other words, it is possible that the observed relationship between taking aspirin and having headaches was a false positive.

1.43 *USA Today* made the mistake of making a cause-and-effect conclusion about the relationship between prayer and blood pressure. This conclusion is not justified because the data were from an observational study. Specifically, they neglected to consider possible confounding variables like lifestyle choices, social networks, and health of the people between the two groups. As a result, people may be led to believe that if they pray more often they will have lower blood pressure. That conclusion is not justified based on this observational study.

1.44 Step 1: The investigators asked if taking aspirin reduces the risk of a heart attack and then in the conclusion, asked to what population the results of this study apply.
Step 2: They collected data from a five-year randomized experiment, including what treatment each doctor received (aspirin or placebo), and whether each doctor had a heart attack or not.
Step 3: They summarized the data from step 2 by categorizing the doctors according to which type of pill they took (aspirin or placebo). They then counted how many doctors in each treatment group had a heart attack. Further data analysis included calculating the "Attacks Per 1000 Doctors" for each group.
Step 4: The investigators proceeded to make conclusions from the data given in Table 1.1. Specifically, they stated, "The results…support the conclusion that taking aspirin does indeed help reduce the risk of having a heart attack. The rate of heart attacks in the group taking aspirin was only about half the rate of heart attacks in the placebo group. In the aspirin group, there were 9.42 heart attacks per 1000 participating doctors, while in the placebo group, there were 17.13 heart attacks per 1000 participants." They then continued to question if there were any other important risk factors or differences between the two groups.
Step 5: The new knowledge is that this study has provided support for the benefit of aspirin in broader populations concerning the reduction of the risk in heart attacks. As a result, millions of people take aspirin with the hope that it will prevent a heart attack.

CHAPTER 2
EXERCISE SOLUTIONS

2.1 **a.** 4
b. A state in the United States.
c. $n = 50$.

2.2 **a.** 2
b. A randomly selected person.
c. $n = 620$.

2.3 **a.** Whole population.
b. Sample

2.4 **a.** Whole population.
b. Sample

2.5 **a.** Population parameter.
b. Sample statistic.
c. Sample statistic.

2.6 **a.** Sample statistic.
b. Population parameter.
c. Sample statistic.

2.7 **a.** Sex and self-reported fastest ever driven speed.
b. Students in a statistics class.
c. The answer may vary. If you think the students represent a larger group of individuals, it is sample data. If interest is only in this group of students, or if you think these students do not represent any larger group, it is population data.

2.8 **a.** $n = 2391$.
b. Individuals aged 65 years or older.
c. Frequency of attending religious services and frequency of praying or reading the bible were related to blood pressure,.
d. Sample data. They used the data to make generalizations about a larger population.

2.9 This is a population summary if we restrict our interest only to the fiscal year 1998. (If we were to use this value to represent errors in other years, it could be considered to be a sample summary.)

2.10 **a.** Treatment used (placebo or aspirin) and whether individual died from heart attack or not.
b. Male physicians between 40 and 84 years old.
c. $n = 22{,}071$.
d. Sample data. The used the data to make generalizations about a larger population.

2.11 **a.** Categorical.
b. Quantitative.
c. Quantitative.
d. Categorical.

2.12 **a.** Quantitative.
b. Categorical.
c. Quantitative.
d. Categorical.

2.13 **a.** Not ordinal. It's categorical but the categories are not ordered.
b. Ordinal. Grades are ordered categories.
c. Not ordinal. It's quantitative.

2.14 **a.** Continuous. All weights are possible within an interval of possibilities (although we can't measure accurately enough to observe all possibilities).
b. Not continuous. The number of text messages must be an integer.
c. Not continuous. The number of coins in a pocket would be an integer.

2.15 **a.** Explanatory variable is score on the final exam; response variable is final course grade.
b. Explanatory variable is sex; response variable is opinion about the death penalty.

2.16 **a.** Not ordinal. It's categorical but the categories are not ordered.
b. Ordinal. The ratings are ordered.
c. Not ordinal. It's quantitative.

2.17 **a.** Not continuous. A student could not miss 4.631 classes for example.
b. Continuous. With an accurate enough measuring instrument, any measurement is possible.
c. Continuous. With an accurate enough time piece, any length of time is possible.

2.18 **a.** Explanatory variable is amount person walks or runs per day; response variable is the performance on the lung test.
b. Explanatory variable is age of the respondent; response variable is feeling about religious importance.

2.19 **a.** Whether a person supports the smoking ban or not is a categorical variable.
b. Gains on verbal and math SATs are quantitative variables.

2.20 The explanatory variable is smoker or not. The response variable is Alzheimer sufferer or not. Both variables are categorical.

2.21 **a.** Sex and pulse rate.
b. Sex is categorical, pulse rate is quantitative.
c. Is there a difference between the mean pulse rates of men and women? The sample mean pulse rate for each sex would be useful.

2.22 This will differ for each student. As an example, suppose a survey question about income only allowed the response categories 1 = under $20,000 and 2= $20,000 to $49,999 and 3= more than $49,999. The income categories are ordered, but so little is known about actual income that the mean response is meaningless.

2.23 This will differ for each student. One example where numerical summaries would make sense for an ordinal variable is the response to the question "What grade do you expect in this class? 1=A, 2=B, 3=C, 4=D, 5=F." The mean numerical response is an expected class GPA.

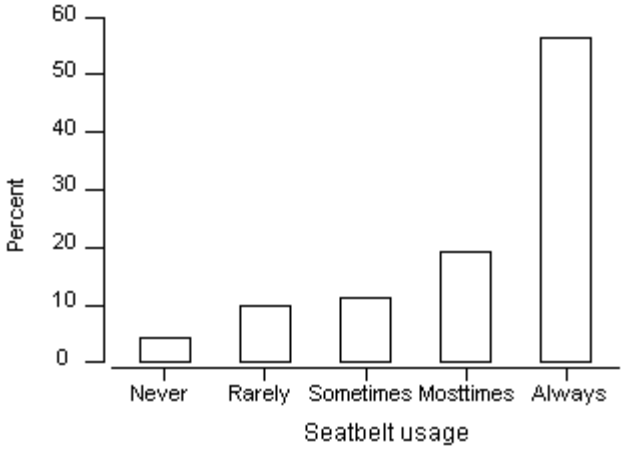
2.24 This will differ for each student.

2.25 **a.** A unit is a person. Dominant hand is a categorical variable and IQ is a quantitative variable. Explanatory variable is dominant hand and response variable is IQ .
b. A unit is a married couple. Eventual divorce status and pet ownership are both categorical variables. Explanatory variable is pet ownership and response variable is eventual divorce status.

2.26 **a.** A unit is a college student. GPA and hours of study each week are both quantitative variables. Explanatory variable is hours of study and response variable is GPA.
b. A unit is a tax-paying individual in the United States. Tax bracket is an ordinal variable and percentage donated to charities is a quantitative variable. Explanatory variable is tax bracket and response variable is percentage donated to charities.

2.27 **a.** 1427/2530 = .564, which is 56.4%.
b. 1 – (1427/2530) = .436, which is 43.6%.
c. Never: 105/2530 = .042 (4.2%); Rarely: 248/2530 = .098 (9.8%); Sometimes: 286/2530 = .113 (11.3%). Most times: 464/2530 = .183 (18.3%); Always: 1427/2530 = .564 (56.4%)
d.

Figure for Exercise 2.27d

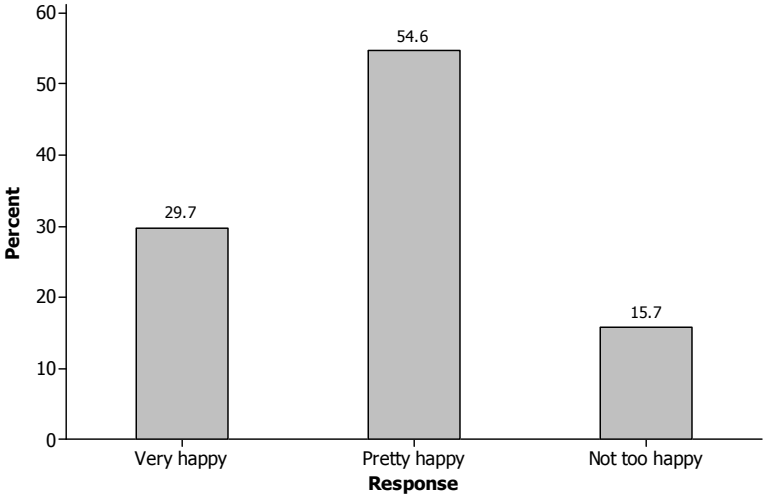


2.28 **a.**

Response	Frequency	Relative frequency
Very happy	599	599/2015 = .297 (29.7%)
Pretty happy	1100	1100/2015 = .546 (54.6%)
Not too happy	316	316/2015= .157 (15.7%)
Total	2015	1 (100%)

b.

Figure for Exercise 2.28b



c. 29.7% + 54.6% = 84.3%

2.29 **a.**

	Preferred use of cell phone		
	To talk	**To text**	**Total**
Women	22 (20.8%)	84 (79.2%)	106 (100%)
Men	34 (41.0%)	49 (59.0%)	83 (100%)

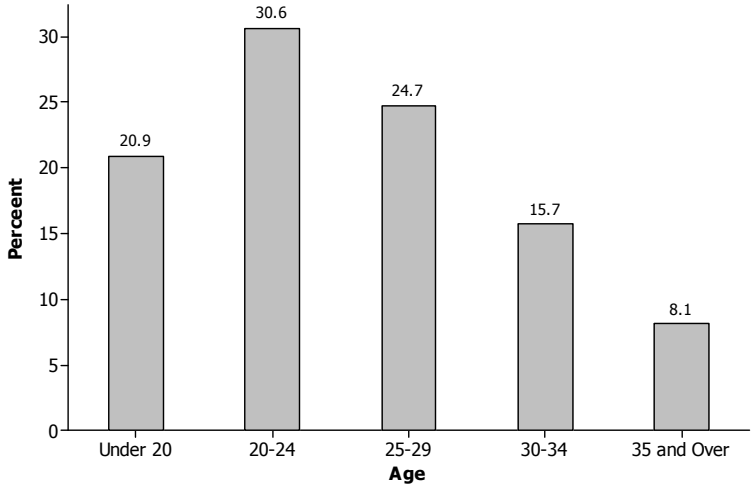
b. Women: 20.8% to talk, 79.2% to text
c. Men: 41.0% to talk, 59.0% to text
d. Women were more likely to say "to text" than men whereas men were more likely to say "to talk."

2.30 **a.** 1700/2470 = .688, or 68.8%
b. 1056/1700 = .621, or 62.1%
c. 300/657 = .457, or 45.7%
d. 41/113 = .363, or 36.3%

2.31 **a.** Explanatory variable is whether a person smoked or not. Response variable is whether they developed Alzheimer's or not.
b. Explanatory variable is political party. Response variable is whether a person voted or not.
c. Explanatory variable is income level. Response variable is whether a person has been subjected to a tax audit or not.

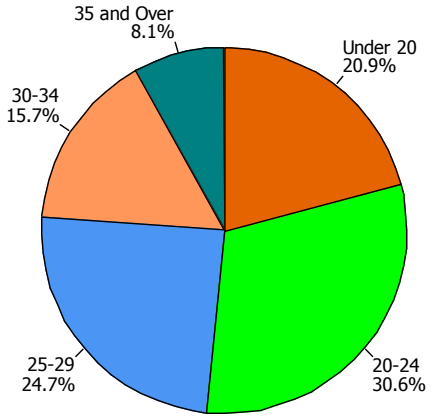
2.32 **a.**

Figure for Exercise 2.32a



b.

Figure for Exercise 2.32b



c. The pie chart may more effectively show that there are three age groups with large percentages, and it may be faster to read these percentages than with the bar chart. One problem, however, is that the age groups are shown in a circular pattern, an unnatural way to view the age. The bar chart gives a better sense of the distribution of ages because the ages are shown along a more natural horizontal number line.

2.33 **a.** The explanatory variable is sex and the response variable is how they feel about their weight.

b.

Feelings About Weight				
Sex	**Overweight**	**About right**	**Underweight**	**Total**
Female	38 (26.6%)	99 (69.2%)	6 (4.2%)	143
Male	18 (23.1%)	35 (44.9%)	25 (32.1%)	78

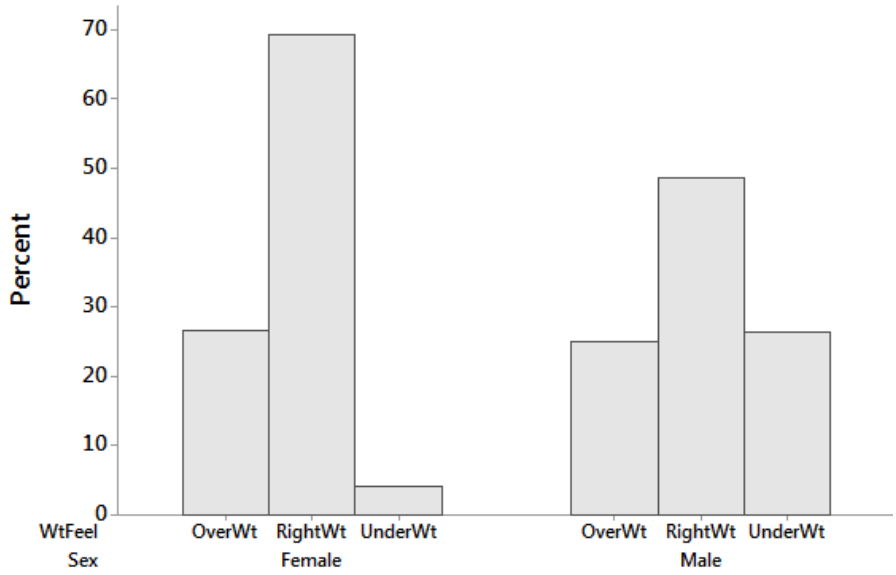
c. Feeling overweight: 38/143 = .266, or 26.6%; right weight: 99/149 = .692, or 69.2%; underweight: 6/149 = .042, or 4.2%.

d. Feeling overweight: 18/78 = .231, or 23.1%; right weight: 35/78 = .449, or 44.9%; underweight: 25/78 = .321 or 32.1%.

e. Males are more likely than females to feel that they are underweight; females are more likely than males to say that their weight is about right.

2.34

Figure for Exercise 2.34



2.35 **a.**

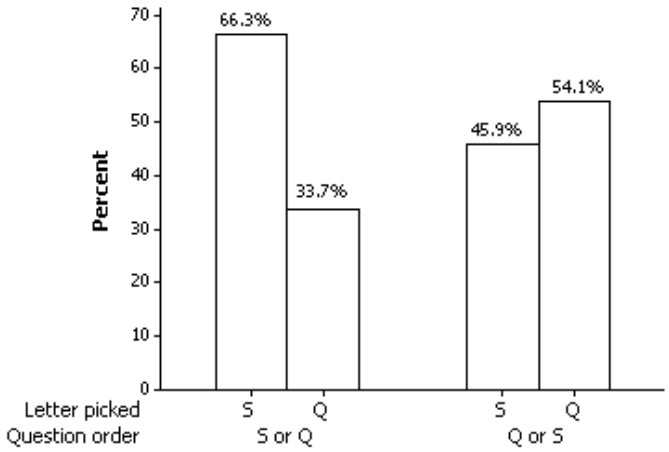
	Picked S	**Picked Q**	**Total**
S listed first	61	31	92
Q listed first	45	53	98
Total	106	84	190

b. Picked S $= (61/92)\times 100\% = 66.3\%$; Picked Q $= (31/92)\times 100\% = 33.7\%$;

c. Picked S $= (45/98)\times 100\% = 45.9\%$; Picked Q $= (53/98)\times 100\% = 54.1\%$;

d.

Figure for Exercise 2.35d



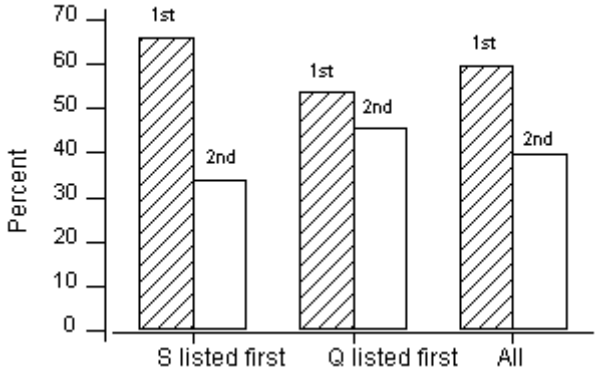
e. Parts (b) and (c) show that the percentage picking S was higher when S was listed first than when Q was listed first. It looks like the letter picked was influenced by the letter listed first.

2.36 **a.** The columns can be labeled "Picked 1st letter" and "Picked 2nd letter." In the "S listed first" row, list the counts in the same order as in the table for Exercise 2.35(a). In the "Q listed first" row, the count for "Q picked" should be in the "Picked 1st letter" column (because Q was the first letter). The table is

	Picked 1st letter	**Picked 2nd letter**	Total
S Listed First	61 (66%)	31 (34%)	92
Q Listed First	53 (54%)	45 (46%)	98
All	114 (60%)	76 (40%)	190

b. A bar chart of percentages picking first and second letters given on each form of the question (row percentages in the table above) along with these percentages for the overall sample follows:

Figure for Exercise 2.36b



c. The variables used in this exercise are more appropriate for illustrating the point of this data set. The question of interest is whether participants might be more likely to pick the fist letter given than the second regardless of whether it was an S or Q. The bar chart given for part (b) illustrates that the fist letter listed was picked more often for both forms.

2.37 **a.** The fastest speed was 150 miles per hour.

b. The slowest speed driven by a male was 55 miles per hour.
c. 1/4 of the females reported having driven at 95 miles per hour or faster. Notice that 95 mph is the *upper quartile* for females. By definition, about 1/4 of the values in a data set are greater than the upper quartile.
d. 1/2 of the females reported having driven 89 mph or faster. Notice that 89 mph is the *median* value.
e. 1/2 of 102 = 51 females have driven 89 mph or faster.

2.38 **a.** The median value is 110 mph for males, compared to 89 mph for females.
b. The spread is about the same for the two sexes. The spread of the extremes is 150− 55 = 95 mph for the males, compared to 130 – 30 = 100 mph for the females. The spread of the quartiles is slightly greater for males (120 – 95 = 25 for males, compared to 95-80 = 15 for females.)

2.39 **a.** The center for the females is at a greater percentage than it is for the males. For females the center looks to be somewhere around 27%. For males, the center looks to be a bit less than 18%.
b. The data are more spread for the females.
c. The greatest two female percentages are set apart from the bulk of the data. The values are about 65% and 72%.

2.40 **a.** Median height = 65 inches.
b. Range = Tallest – Shortest = 71 – 59 = 12 inches.
c. The interval from 59 to 63.5 inches contains the shortest 1/4 of the women.
This interval is from the minimum to the lower quartile.
d. The interval from 63.5 to 67.5 inches contains the middle 1/2 of the women.
This is an interval from the lower quartile to the upper quartile.

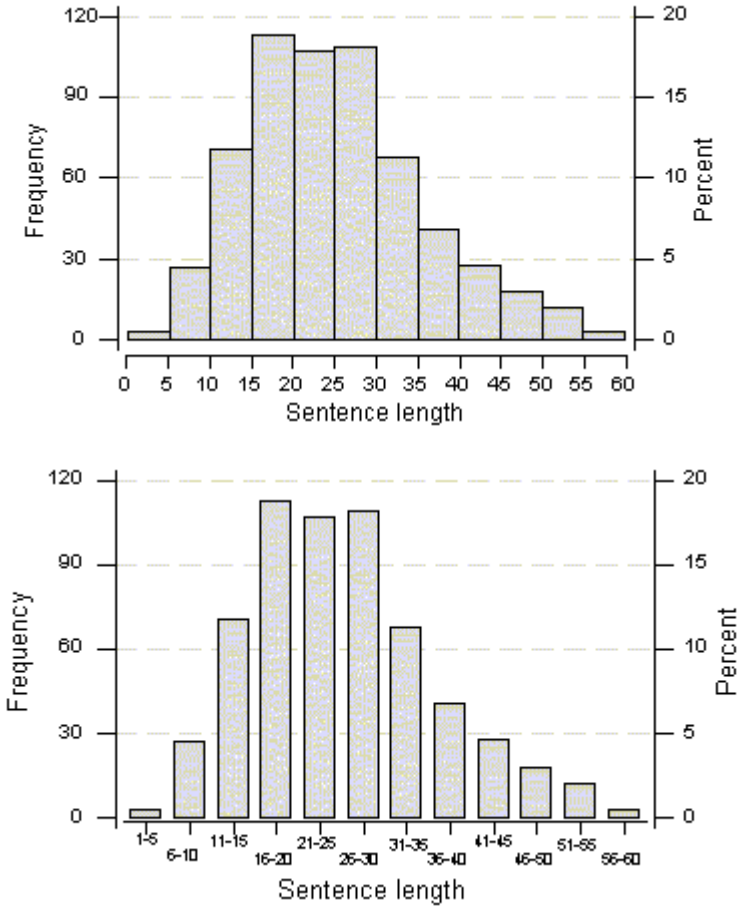
2.41 **a.** The median value, 65 inches, describes the location.
b. The interval described by the extremes, 59 to 71 inches, describes spread. We might also describe spread using the interval 63.5 to 67.5, the spread of the middle 50% of the data.

2.42 **a.** The dataset is skewed to the right (it stretches in that direction).
b. The value 13 looks to be an outlier. It is separate from the bulk of the data.
c. 2 ear pierces was the most reported value. About 44 or so women said they had this many ear pierces.
d. About 32 or so women said they had 4 ear pierces.

2.43 **a.** The dataset looks approximately symmetric and bell-shaped.
b. There are no noticeable outliers.
c. The most frequently reported value for sleep was 7 hours.
d. Roughly 14 or so students said they slept 8 hours the previous night.

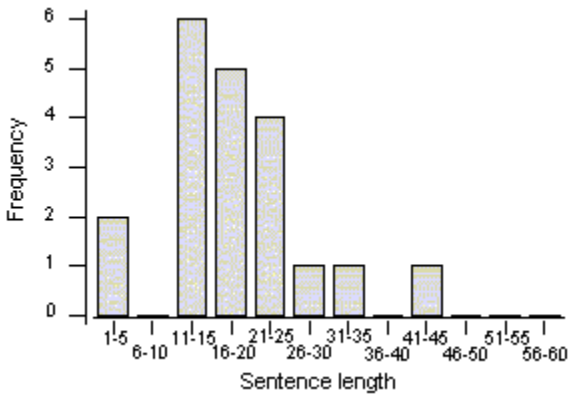
2.44 **a.** Two slightly different versions of the histogram are shown below. In the first, the bars touch each other; in the second, the bars are separated and labeled with the category limits. The first histogram looks a little nicer but there could be confusion about the exact endpoints of intervals. Notice that we have shown a frequency axis on the left and the corresponding relative frequency (percentage) axis on the right.

Figures for Exercise 2.44a



b. A majority of the sentences have between 16 and 30 words. The percentage of sentences with between 16 and 30 words is (113+107+109)/600 ≈55%. With regard to shape, the data are skewed to the right.
c. A stem-and-leaf plot presents all individual values, but we do not have the data in this form. It has already been tallied for specified intervals.
d. The lead-in to the definition of statistics given on page 1 might cause some confusion. If the lead-in (beginning "A more complete definition, and ...") and the definition are counted together as one sentence, that sentence has 42 words. Because the definition is set apart in a box, many may divide these 42 words into two sentences - one with 17 words and one with 25 (the definition itself). The histogram shown below is for the first possibility (one sentence of 42 words).

Figure for Exercise 2.44d



The histogram of number of words in a sentence in the first twenty sentences of Chapter 1 is also skewed to the right. In general, however, the number of words per sentence is less than that in the *Shorter History of England.*

2.45 **a.** The dataset looks approximately symmetric and bell-shaped.
b. The highest temperature is 92°F.
c. The lowest temperature is 64°F.
d. 5/20 = .25, which is 25%.

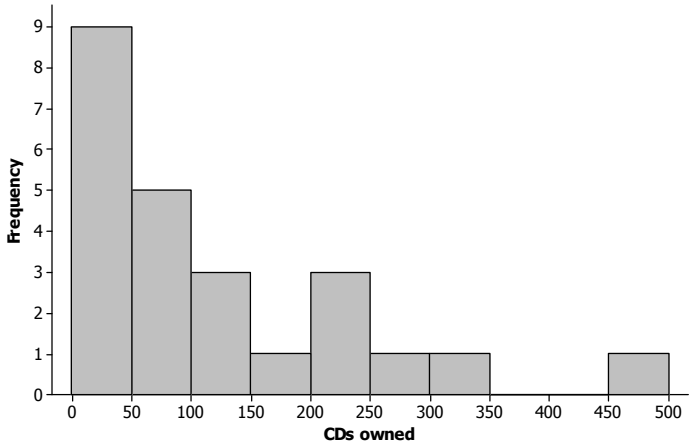
2.46 **a.** The following stem-and-leaf uses "hundreds" for stems (row labels), "tens" for leaves within the rows and truncates the "ones" digit. Each "hundreds" is represented by two separate stems, one for leaves 0-4 and one for leaves 5-9.

Figure for Exercise 2.46a

```
|0| 001222233
|0| 55569
|1| 002
|1| 5
|2| 002
|2| 5
|3| 0
|3|
|4|
|4| 5
```

b. In the following histogram, values tied with the lower value of an interval are counted into that interval. For example, 50 is counted into the interval that spans from 50 to 100.

Figure for Exercise 2.46b



c. The data are skewed to the right with a possible outlier (at 450).

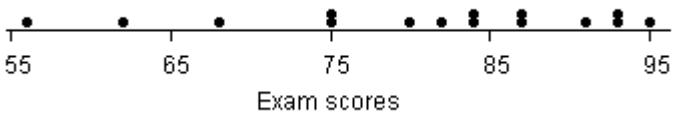
2.47 **a.** The figure below uses separate stems for last digits 0-4 and 5-9. That's not imperative, although doing so give more detail of the shape of the data.

Figure for Exercise 2.47a

```
|5| 6
|6| 2
|6| 8
|7|
|7| 55
|8| 02344
|8| 77
|9| 13
|9| 5
```

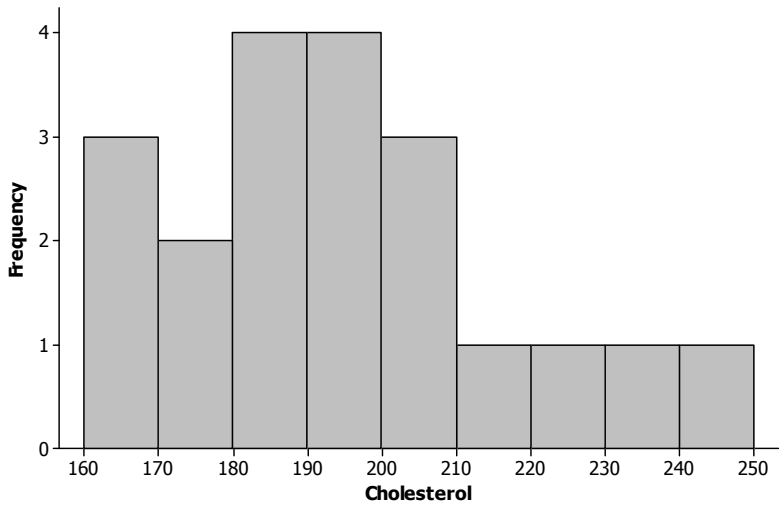
b.

Figure for Exercise 2.47b



2.48 **a.** The following histogram uses nine intervals. Others are possible.

Figure for Exercise 2.48a



b. In the following stem-and-leaf, the first two digits of the values are used for stems (row labels).

Figure for Exercise 2.48b

```
|16| 046
|17| 88
|18| 2248
|19| 6888
|20| 046
|21| 2
|22| 2
|23| 0
|24| 2
```

c. There are no obvious outliers.
d. The shape is difficulty to judge – perhaps a slight skew to the right.

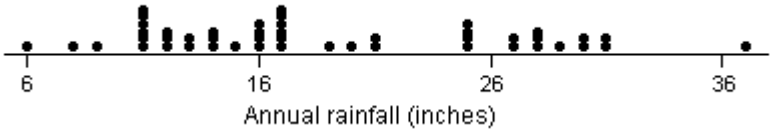
2.49 **a.** In the figure shown here, two stems have been used for each possible "tens" place in the number (values under 10 are an exception because the lowest value is about 6 inches). We rounded the 1995 total of 24.5 inches up to 25.

Figure for Exercise 2.49a

```
|0| 689
|1| 11111122233444
|1| 566667777779
|2| 011
|2| 5555778889
|3| 0011
|3| 7
```

b.

Figure for Exercise 2.49b



c. The data are skewed (but only slightly) to the right.

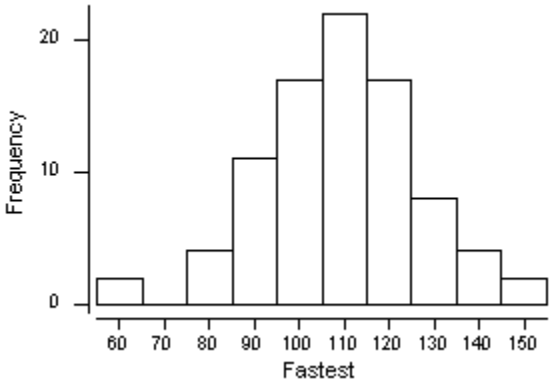
2.50 *Note*: The data for males are in the dataset **pennstate1M** on the companion website.
a.

Figure for Exercise 2.50a

```
| 5| 5
| 6| 0
| 7|
| 8| 00005555
| 9| 0000024555555
|10| 000000000012555555559
|11| 0000000000002555555
|12| 00000000004555555
|13| 00
|14| 00005
|15| 0
```

b. The histogram shown here was done with Minitab. Notice that the midpoints of intervals are shown under the bars.

Figure for Exercise 2.50b

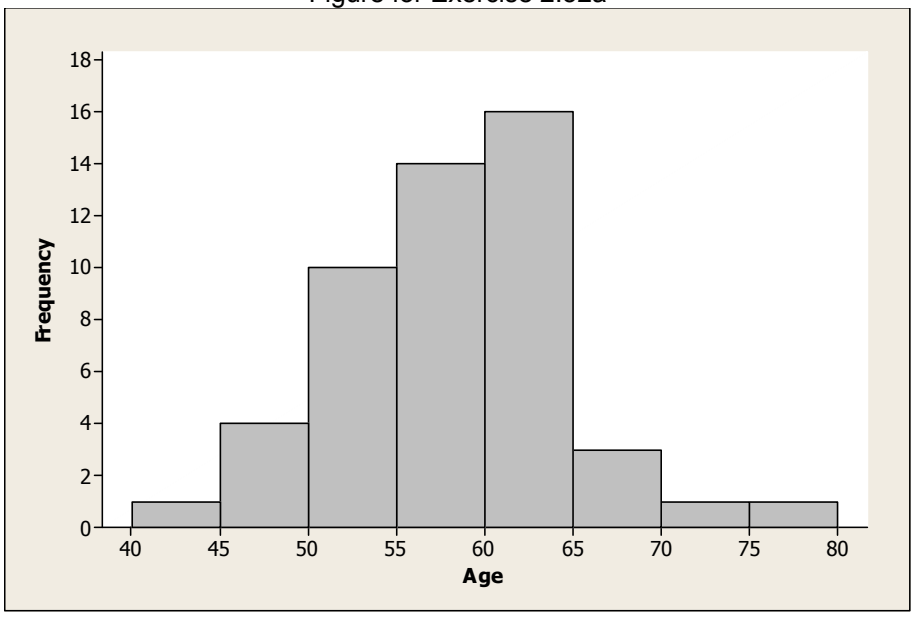


c. The histogram may provide the best view of the overall shape and characteristics of the distribution. The information given in the stem-and-leaf plot is equivalent to that given in the histogram, but is somewhat harder to read. An advantage of the stem-and-leaf is that it's possible to determine individual values. The dotplot gives much information about individual values as well as the range and general location. It is, however, more difficult to judge the shape of the distribution with a dotplot.
d. These data are symmetric in shape (and there may be two outliers).

2.51 Yes, a stem-and-leaf plot provides sufficient information to determine whether a dataset contains an outlier. Because all individual values are shown, it is possible to see whether are any values are inconsistent with the bulk of the data.

2.52 **a.** The answer will vary due to the flexibility possible for deciding on the endpoints of intervals. The histogram shown here is based on 5-year age groups and the age under the left edge of a bar is included in that interval while the age under the right edge is not.

Figure for Exercise 2.52a



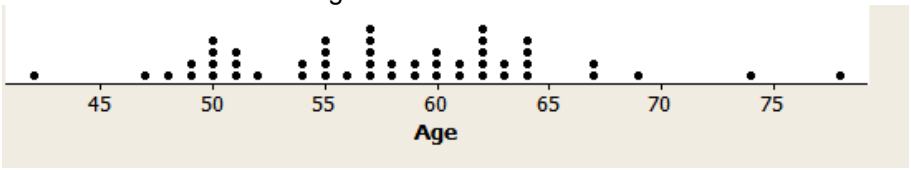
b. Two stems should be used per decade of ages because there should be 6 to 15 stem values. [Note: If only one stem value is used per decade, there would be only 5 values, and if 5 were used per decade there would be 22 (because only 4 would be needed for the 30s and 3 for the 70s).] Notice that if this stem-and-leaf plot were turned on its side, it would have the same shape as the histogram shown for part (a).

Figure for Exercise 2.52b

```
|4| 2
|4| 7899
|5| 0000111244
|5| 55556777778899
|6| 0001122222334444
|6| 779
|7| 4
|7| 8
```

c.

Figure for Exercise 2.52c



d. The CEO ages may be skewed slightly to the left although this is not a pronounced skew.. It may be fair to say that the data are roughly bell-shaped.

e. There do not seem to be any obvious outliers although there is a bit of a gap between the oldest two CEOs and the third oldest and also a small gap between the youngest and the others.
f. This is a conjecture, but it seems more likely that an outlier would occur in the salaries. A person's salary can become very high, and a company might conceivably pay an extraordinary salary to a successful executive.

2.53 Skewed to the left. Along a number line, values stretch more toward the left (small values).

2.54 Answers will differ. One approach is to combine two distinctly different groups into the same dataset. As an example, a histogram of the heights of a sample that includes equal numbers of adult women and six-year old girls will be bimodal.

2.55 **a.** Histogram is better than a boxplot for evaluating shape.
b. A boxplot is useful for identifying outliers, evaluating spread, and for comparing groups.

2.56 The answers will differ for each student.
a. You may be most interested in knowing the average value because it would provide some information about what kind of salary to expect. You may also like to know the spread because the average value would be less important if the annual salary of employees varied widely.
b. Each summary has interest here. The maximum would give information about whether an A is possible with each instructor. For a generally average student, the average might have the most interest. The spread would give information about whether most students performed about the same or whether there was great variability among students.
c. The average may be the most informative about personal life expectancy, although the maximum and the spread would also be useful and interesting general information.

2.57 Generally, females tended to have higher tip percentages. The median is clearly greater for females. The data for the females also shows greater spread than the data for the males.

2.58 **a.** Median = 70. Ordered list of data is 67, 68, 69, 70, 72, 73, 74; median is middle value in this ordered list.
b. Mean = 70.43

2.59 **a.** Mean = 74.33; median = 74.
b. Mean = 25; median = 7.
c. Mean = 27.5; median = 30.

2.60 100 is a large value compared to the rest of the values; it causes the mean to increase, while not affecting the median.

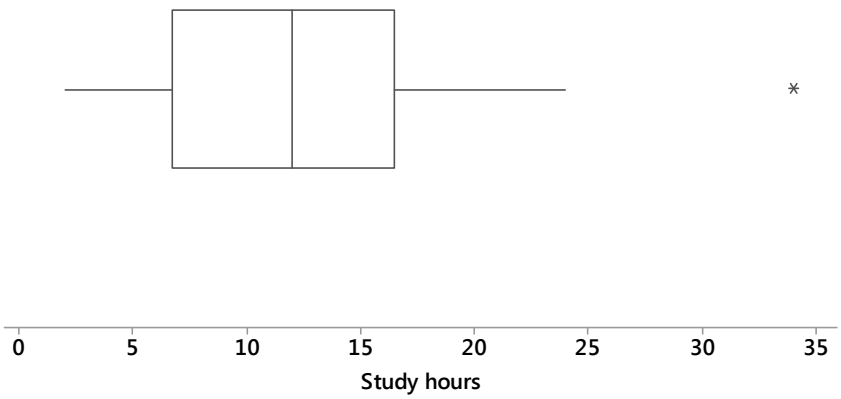
2.61 **a.** 225 –123 = 102 lbs
b. 190 –155 = 45 lbs
c. 50%

2.62 **a.** 12 letters.
b. 13 letters.
c. The IQR for males is 17–10 = 7 while for females it is 15–10 = 5. The IQR is larger for males.
d. 23 – 6 = 17 letters.
e. 23 – 6 = 17 letters.

2.63 **a.** Min = 2, Q_1 = 7, Median = 12, Q_3 = 16, Max = 34
The ordered data follow. A vertical bar indicates the location of the median. Values of Q_1 and Q_3 are underlined and in bold, They are the medians of the lower and upper halves of the data, respectively.
2 3 4 6 6 **7** 8 9 10 11 12 | 12 13 14 15 15 **16** 18 21 22 24 34
b. The boxplot is below. The lines extend at most 1.5 IQR from the ends of the box, which is 1.5(16 – 7) = 13.5, but stop at the Min and Max if those are reached first. Therefore, the line at the lower end stops at 2,

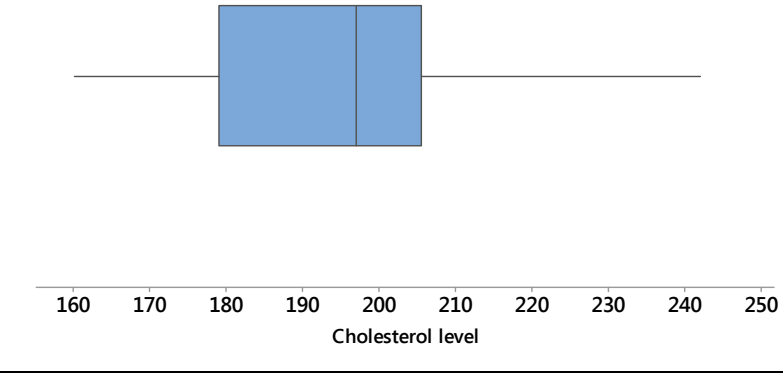
but the line at the upper end extends to (16 + 13.5) = 29.5. Note that the value 34 is marked as an outlier. It exceeds the upper boundary for an outlier, which is Q3 + 1.5 IQR = 16 + 1.5 (16 − 7) = 29.5.

Figure for Exercise 2.63b



2.64 **a.** Min = 160, Q_1 = 180, Median = 197, Q_3 = 205, Max = 242
The ordered data follow. Vertical bars indicates the locations of Q_1, the median, and Q_3, respectively.
160 164 166 178 178 | 182 182 184 188 196 |
198 198 198 200 204 | 206 212 222 230 242
b. The boxplot is below. The lines extend at most 1.5 IQR from the ends of the box, which is 1.5(205 – 180) = 37.5, but stop at the Min and Max if those are reached first. Therefore, the line at the lower end stops at 160 instead of extending to Q1 – 37.5 = 180 – 37.5 = 142.5, and the line at the upper end stops at 242 instead of extending to (205+37.5) = 242.5.

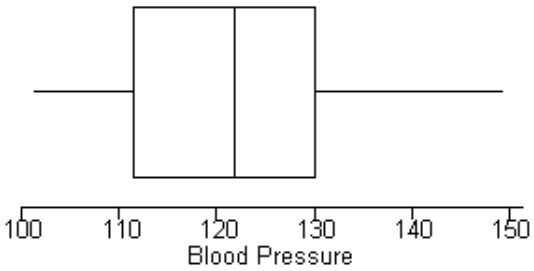
Figure for Exercise 2.64b



2.65 **a.** Min = 109, Q_1 = 180.75, Median = 186, Q_3 = 199.0, Max = 214.0
The data in order are 109.0, 178.5, 183.0, 185.0, 186.0, 188.5, 194.5, 203.5, 214.0.
Median is the middle value (= 186). Lower quartile is median of 109.0, 178.5, 183.0, 185, so equals (178.5 + 183).2 = 180.75. Upper quartile is median of 188.5, 194.5, 203.5, 214, so equals (194.5 + 203.5)/2 = 199.0
b. 109.0 is an outlier. It's below the lower outlier boundary = $Q_1 - 1.5\,IQR$ = 180.75 – 1.5 (199 – 180.75) = 153.375.
c. The member who weighed 109.

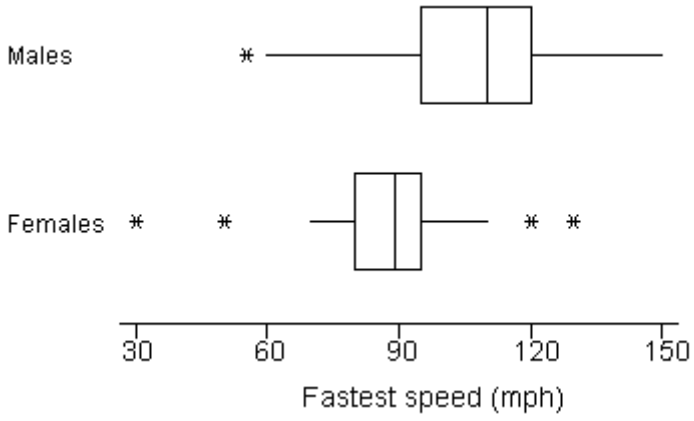
2.66 **a.** (122+123)/2 = 122.5.
b. Q_1 = 114 and Q_3 = 129.5.
c. IQR = $Q_3 - Q_1$ = 129.5–114 = 15.5.
d. 1.5×IQR = 23.25. A number will be considered an outlier if it is either below 114–23.25 = 90.75 or above 129.5+ 23.25 = 152.75. No values fit this criterion, so there are no outliers.
e.

Figure for Exercise 2.66e



2.67 The boxplots below show that, on average, the fastest speeds ever driven by males tend to be higher than the fastest speeds ever driven by females. It is also seen, if outliers are ignored, that the spread is greater for males than it is for females. A horizontal axis has been used for fastest speeds here, but a vertical axis would be equally appropriate.

Figure for Exercise 2.67



2.68 **a.** The amount of exercise per week is similar for men and women. The dotplot follows.