Section 2.1 Different Types of Data

2.1 Categorical/quantitative difference

a) Categorical variables are those in which observations belong to one of a set of categories, whereas quantitative variables are those on which observations are numerical.

b) An example of a categorical variable is religion. An example of a quantitative variable is temperature.

2.2 Common types of cancer in 2012

The variable summarized is categorical. The variable is type of cancer an individual can be affected by. There are six types of categories, such as, lung cancer, breast cancer, colorectal cancer, prostate cancer, stomach cancer, and other cancer types. These types are the categories.

2.3 Classify the variable type

a) quantitative
b) categorical
c) categorical
d) quantitative

2.4 Categorical or quantitative?

a) categorical
b) quantitative
c) categorical
d) quantitative

2.5 Discrete/continuous

a) A discrete variable is a quantitative variable for which the possible values are separate values such as 0, 1, 2, …. A continuous variable is a quantitative variable for which the possible values form an interval.

b) Example of a discrete variable: the number of children in a family (a given family can't have 2.43 children). Example of a continuous variable: temperature (we can have a temperature of 48.659).

2.6 Discrete or continuous?

a) continuous
b) discrete
c) continuous
d) discrete

2.7 Discrete or continuous 2

a) continuous
b) discrete
c) discrete
d) continuous

2.8 Number of children

a) The variable, number of children, is quantitative.

b) The variable, number of children, is discrete.

c)

No. children	0	1	2	3	4	5	6	7	8+
Count	521	323	524	344	160	77	30	19	22
Proportion	0.258	0.160	0.259	0.170	0.079	0.038	0.015	0.009	0.011
Percentage	25.8	16.0	25.9	17.0	7.9	3.8	1.5	0.9	1.1

2.9 Fatal Shark Attacks

a)

Location	Florida	Hawaii	California	Australia	South Africa
Count	2	2	4	15	13
Proportion	0.032	0.032	0.063	0.238	0.206
Percentage	3.2	3.2	6.3	23.8	20.6

Location	Reunion Island	Brazil	Bahamas	Other
Count	6	4	6	11
Proportion	0.095	0.063	0.095	0.175
Percentage	9.5	6.3	9.5	17.5

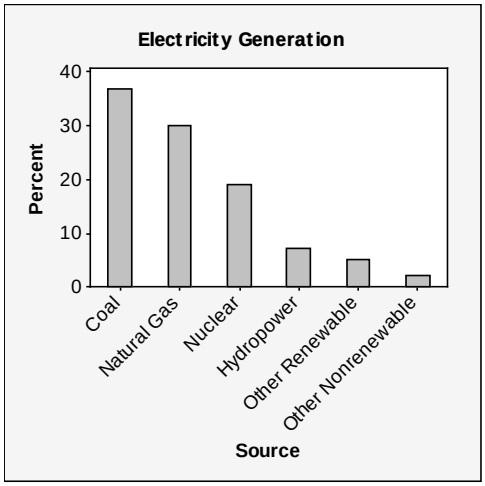
b) Australia is the modal category.

c) The regions with most frequent fatal shark attacks are Australia and South Africa.

Section 2.2 Graphical Summaries of Data

2.10 Generating Electricity

a)



b) Sketching a bar chart would be easier. Sketching the precise areas corresponding to the percentages is more challenging in a pie chart.

c) It is straightforward to judge the relative sizes when comparing the bars corresponding to the percentages.

d) Coal is the modal category.

2.11 What do alligators eat?

a) Primary food choice is categorical.

b) The modal category is "fish."

c) Approximately 43% of alligators ate fish as their primary food choice.

d) This is an example of a Pareto chart, a chart that is organized from most to least frequent choice.

2.12 Weather stations

a) The slices of the pie portray categories of a variable (i.e., regions).

b) The first number is the frequency, the number of weather stations in a given region. The second number is the percentage of all weather stations that are in this region.

c) It is easier to identify the modal category using a bar graph than using a pie chart because we can more easily compare the heights of bars than the slices of a piece of pie. For example, in this case, the slices for Midwest and West look very similar in size, but it would be clear from a bar graph that West was taller in height than Midwest.

2.13 France is most popular holiday spot

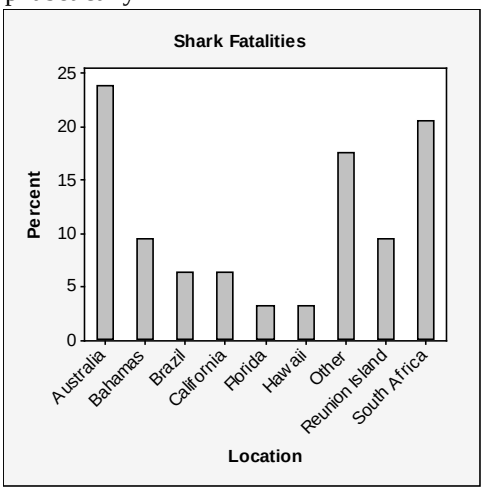
a) Country visited is categorical.

b) A Pareto chart would make more sense because it allows the viewer to easily locate the categories with the highest and lowest frequencies.

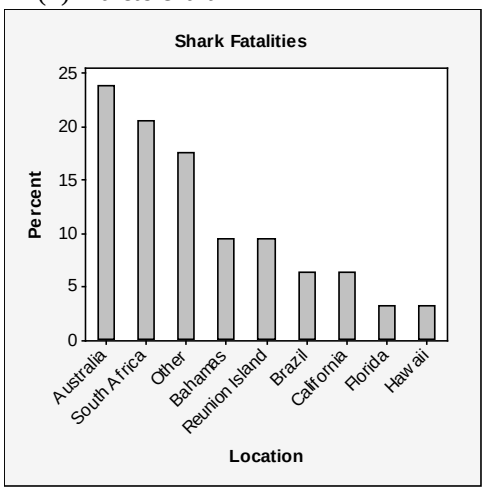
c) A dot plot or stem-and-leaf plot do not make sense because the data are categorical; these two types of plots are used with quantitative data (and also with data that have relatively few observations).

2.14 Pareto chart for fatal shark attacks

(i) Alphabetically



(ii) Pareto chart



With a Pareto chart, it is straightforward to identify the few regions with the largest number of fatal shark attacks.

2.15 Sugar dot plot

a) The minimum sugar value is zero grams, and the maximum is 18 grams.

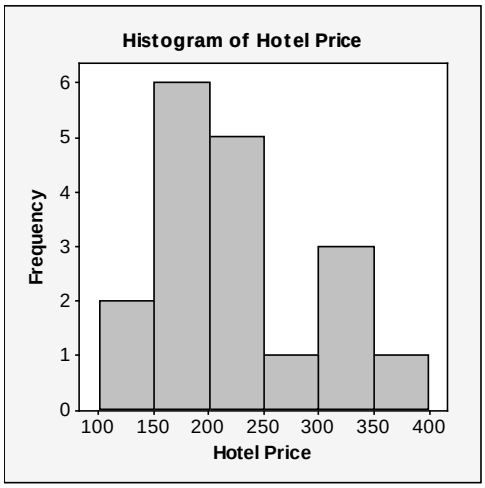
b) The sugar outcome that occurs most frequently is called the mode. For this data set there are five modes: three, four, eleven, twelve and fourteen grams.

2.16 Spring break hotel prices

a) 1 | 24677999
2 | 133445
3 | 1338

b) 1 | 24
1 | 677999
2 | 13344
2 | 5
3 | 133
3 | 8
The plot with split stems gives a clearer picture of the shape of the distribution.

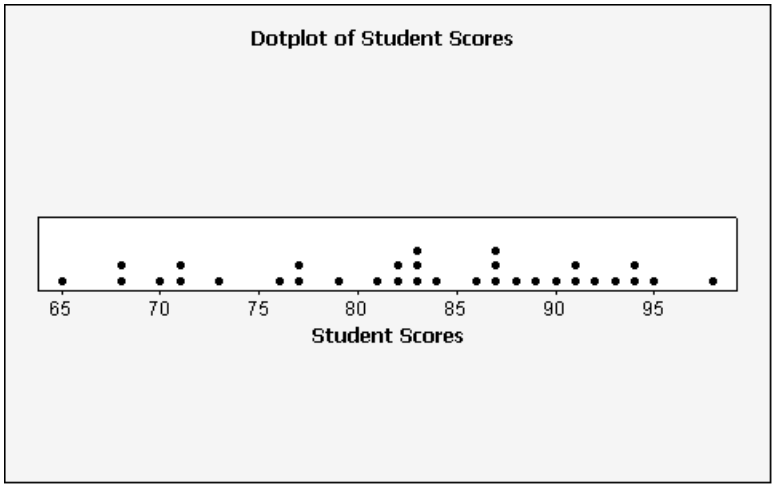
c) Most hotels charge between $150 and $250 per night, with a few charging more. The distribution of prices is right-skewed.



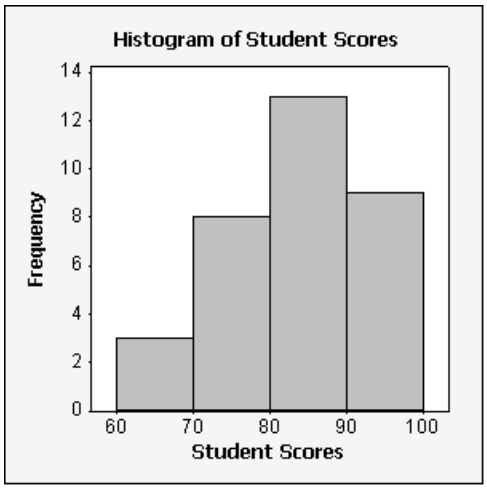
2.17 Student scores

a) There are 33 validated courses. The minimum score is 65 and the maximum score is 98.

b)



c)



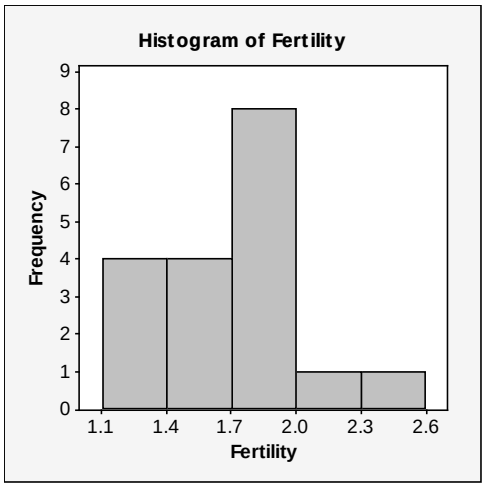
2.18 Fertility rates

a) 1 | 3333445677778899
2 | 04
A disadvantage of this plot is that it is too compact making it difficult to visualize where the data fall.

b) 1 | 333344
1 | 5677778899
2 | 04

2.18 (continued)

c)



2.19 Split Stems

a) smallest 0 g, largest 18 g
b) 10 g, 11 g, 11 g
c) Six cereals have less than 5 grams of sugar with 0 g, 1 g, 3 g, 3 g, 4 g, and 4 g.

2.20 Histogram for sugar

a) –1 to 1,1 to 3, 3 to 5, 5 to 7, 7 to 9, 9 to 11, 11 to 13, 13 to 15, 15 to 17, and 17 to 19
b) The distribution is bimodal; child cereals, on average, have more sugar than adult cereals have.
c) The dot and stem-and-leaf plots allow us to see all the individual data points.
d) The relative differences among bars would remain the same.

2.21 Shape of the histogram

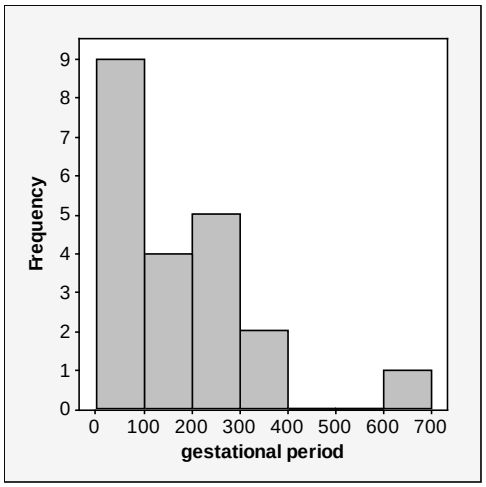
a) Price of a certain model of smartwatch – symmetric because most prices would fall in the middle with some prices being higher and some lower.
b) Amount of time students used to take an exam in your school – skewed to the left because only some students are most likely to leave early, more of them stay late, and many of them stay until the end of the allowed time.
c) The grade point average (GPA) in your academic program this year-symmetric because most GPAs would be in the middle with some higher and some lower.
d) The salary of all the employees in a company – skewed to the right because most of employees have salaries in the lowest range, few salaries are in the middle range and very few salaries are in the highest range.

2.22 More shapes of histograms

a) The winner's score in a basketball game during an NBA season – symmetric because most game winner's scores would fall in the middle, with some scores being higher and some lower.
b) The distance from home to school for students in a specific school – skewed to the right because most of students' homes are very close to the school, a few are a little bit far in the middle range and very few are very far from the school.
c) The number of attempts a young adult needs to pass a driving license test – skewed to the right because most of young adults would pass their driving license test within few attempts, few of them will need more attempts and very few young adults will need many attempts to pass their driving license test.
d) The number of times an individual request a password reset for the forgotten password of his or her main email account – skewed to the right because most of individuals are most likely to request a few number of password reset, few of them will request more password reset and very few will request many password reset.

2.23 Gestational Period

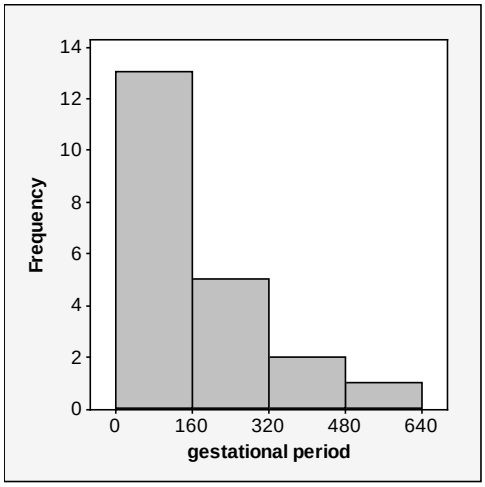
a)

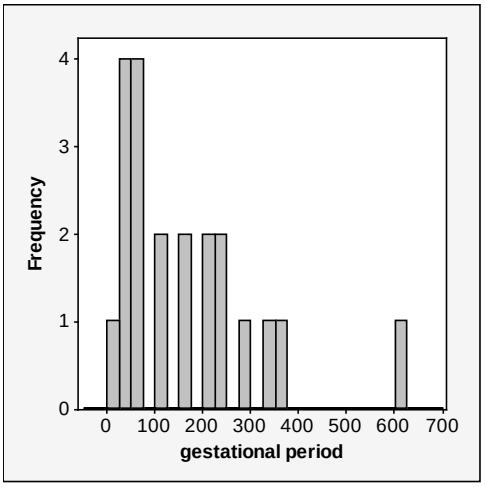


b) The elephant, with a gestational period of 624 days, is unusual.

c) The distribution is right-skewed.

d) Neither of the two histograms accurately summarizes the distribution. The one with 4 intervals is too coarse, the one with 30 intervals too fine.





2.24 How often do students read the newspaper?

a) This is a discrete variable because the value for each person would be a whole number. One could not read a newspaper 5.76 times per week, for example.

b) (i) The minimum response is zero.
(ii) The maximum response is nine.
(iii) Two students did not read the newspaper at all.
(iv) The mode is three.

c) This distribution is unimodal and somewhat skewed to the right.

2.25 Blossom widths

a) The distribution is slightly right-skewed (or roughly symmetric). Most blossoms have a width between 3.2 and 3.6 in. There is one blossom with an unusual small width for that species of less than 2.4 in.

b) The distribution is left-skewed. Most blossoms have a width between 2.8 and 3.2 in.

c) $\frac{6+15+24}{50} = 0.90$ or 90%

d) No. We don't know how many blossoms in the interval from 2.8 to 3.2 in. are actually wider than 3 in.

2.26 Central Park temperatures

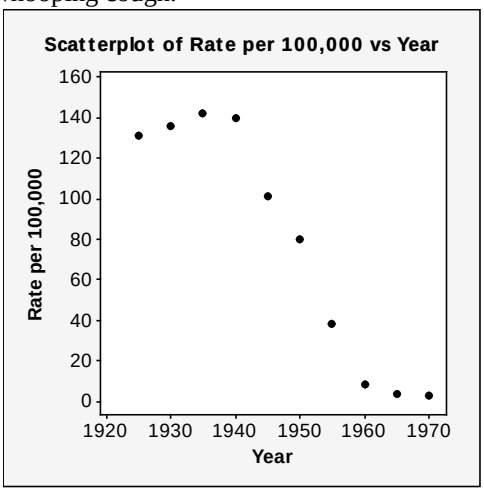
a) The distribution is somewhat skewed to the left.

b) A time plot connects the data points over time to show time trends.

c) A histogram shows the number of observations at each level more easily than does the time plot. We also can see the shape of the distribution from the histogram but not from the time plot.

2.27 Is whooping cough close to being eradicated?

a) One can see in the time plot below that after an initial slight increase, there was a sharp and steady decrease in incidence of whooping cough starting around 1940. The decrease leveled off starting around 1960. These data suggest that the whooping cough vaccination was proving effective in reducing the incidence of whooping cough.

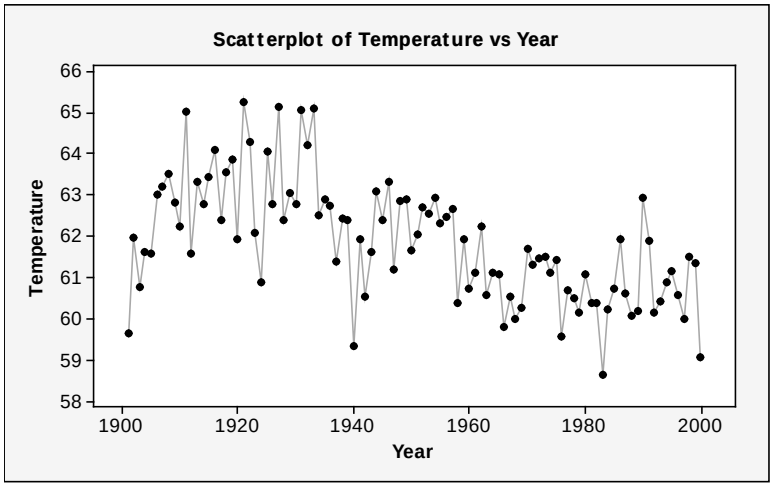


b) The incidence rate stayed low until about 2000, after which a sharp increase can be observed. No, the United States is not close to eradicating whooping cough. Potential reasons for this include fewer people deciding to get vaccinated and less efficient vaccinations.

c) A histogram would not address this question because it does not show the rates for each year; we would not be able to see changes over time.

2.28 Warming in Newnan, GA?

Overall, the time plot (below) does seem to show a decrease in temperature over time.



Section 2.3 Measuring the Center of Quantitative Data

2.29 Median versus mean

a) Median (The distribution would be right-skewed.)
b) Median (The distribution would be left-skewed.)
c) Mean (The distribution would be symmetric.)

2.30 More median versus mean

a) Mean (The distribution would be symmetric.)
b) Median (The distribution would be right-skewed.)
c) Median (The distribution would be right-skewed.)

2.31 More on CO_2 emissions

a) Mean: $\bar{x} = \frac{\sum x}{n} = \frac{8.0+5.3+1.8+1.7+1.2+0.8+0.6+0.5+0.4+0.4}{10} = \frac{20.7}{10} = 2.07$

Median: Find the middle value: $\frac{n+1}{2} = \frac{10+1}{2} = 5\frac{1}{2}$ th position

0.4, 0.4, 0.5, 0.6, 0.8, 1.2, 1.7, 1.8, 5.3, 8.0

The median is $\frac{0.8+1.2}{2} = 1$.

b) Comparing absolute emission values for nations with different population sizes might be misleading because nations with larger populations tend to have larger total emissions. When viewed per capita, a different picture might emerge.

2.32 Resistance to an outlier

a) The median for all three data sets is ten. The values for all three sets of observations are already arranged in numerical order, and the middle number for each is 10.

b) Set 1: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+12}{5} = \frac{50}{5} = 10$

Set 2: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+100}{5} = \frac{138}{5} = 27.6$

Set 3: $\bar{x} = \frac{\sum x}{n} = \frac{8+9+10+11+1000}{5} = \frac{1038}{5} = 207.6$

c) As the highest value becomes more and more of an extreme outlier, the median is unaffected, whereas the mean increases as the outlier becomes more extreme.

2.33 Weekly earnings and gender

For both males and females weekly earning distributions, the mean is much larger than the median which suggests a right-skewed shape for both of them.

2.34 Labor dispute

Management would want to use the mean because it would be skewed right by the outliers – the few members of management who make a whole lot of money. The mean income would be higher because of the outliers. The workers would prefer the median because it is not affected by the large outliers. It is a more accurate measure of the actual typical income.

2.35 Cereal sodium

The moderate skewness to the left causes the mean to be lower than the median.

2.36 Center of plots

a) The mean and median would be the same for the dot plots to the middle and to the right because the distributions are symmetric.
b) The distribution to the left is skewed to the right, and the mean would be higher than the median would. The mean would be pulled toward the higher, atypical values.

2.37 Public transportation – center

a) The mean is 2, the median is 0, and the mode is 0. Thus, the average score is 2, the middle score is 0 (indicating that the mean is skewed by outliers), and the most common score also is 0.

Mean: $\bar{x} = \frac{\sum x}{n} = \frac{0+0+4+0+0+0+10+0+6+0}{10} = \frac{20}{10} = 2$

Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, 10

Mode: the most common score is zero.

b) Now the mean is 10, but the median is still 0.

Mean: $\bar{x} = \frac{\sum x}{n} = \frac{0+0+4+0+0+0+10+0+6+90}{10} = \frac{110}{10} = 10$

Median: middle score of 0, 0, 0, 0, 0, 0, 0, 4, 6, 10, 90

The median is not affected by the magnitude of the highest score, the outlier. Because there are so many zeros, even though we've added one score, the median remains zero. The mean, however, is affected by the magnitude of this new score, an extreme outlier.

2.38 Public transportation – outlier

a) The mean versus median applet confirms that the median is not affected by the magnitude of the highest score. Because there are so many zeros, even though we've added one score, the median remains zero. The mean, however, is affected by the magnitude of this new score, an extreme outlier.

b) The applet demonstrates that the outlier has a weaker effect when there are more scores near the original mean.

2.39 Sale price of houses

The mean of the sale price of houses in the United States is much larger than the median, which suggest a right-skewed distribution. There are a few expensive houses with a high sales price.

2.40 More baseball salaries

Answers will vary.

2.41 European fertility

a) The median fertility rate is 1.7. Thus, about half of the countries listed have mean fertility rates at or below 1.7 with the remaining countries having fertility rates above 1.7.

b) The mean of the fertility rates is 1.65.

c) Since the population of adult women can vary greatly among the countries, it is necessary to calculate an overall fertility rate for the country in order to make comparisons. This rate is found by calculating the mean number of children per adult woman. The mean for a variable need not be one of the possible values for the variable. Although the number of children born to each adult woman is a whole number, the mean number of children born per adult woman need not be a whole number. For example, the mean number of children per adult woman is considerably higher in Mexico than in Canada.

2.42 Dining out

a) The numbers in the 285th and 286th positions of the sorted data set are 1. The median is an average of these two values (i.e., 1).

b) Mean: $\bar{x} = \frac{\sum x_i}{n} = \frac{84(0)+290(1)+100(2)+46(3)+30(4)+13(5)+5(6)+2(7)}{570} \approx 1.5$

c) Since 290 respondents gave an answer of 1, the 285th and 286th positions of the sorted data set from smallest to largest are 1 and so is the median. The median is still 1. However, the value of the mean changes:

Mean: $\bar{x} = \frac{\sum x_i}{n} = \frac{290(1)+100(2)+46(3)+30(4)+13(5)+5(6)+86(7)}{570} \approx 2.54$

2.43 Marriage statistics for 20–24-year-olds

a) Women: The mean is 0.274, the median is 0.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{7350(0) + 2587(1) + 80(2)}{10,017} = \frac{2747}{10,017} = 0.274$$

The median is the middle score. With 10,017 scores, the median is the score in the 5009th position. Thus, the median is 0.

Men: The mean is 0.161, the median is 0.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{8418(0) + 1594(1) + 10(2)}{10,022} = \frac{1614}{10,022} = 0.161$$

The median is the middle score. With 10,022 scores, the median is the score between the 5011th and 5012th positions. Thus, the median is 0.

b) Using the medians, it seems that there is no difference. Using the mean, in this age group, women have, on average, been married more often.

2.44 Knowing homicide victims

a) The mean is 0.16.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3944(0) + 279(1) + 97(2) + 40(3) + 23(4.5)}{4383} = \frac{696.5}{4383} = 0.16$$

b) The median is the middle score. With 4383 scores, the median is the score in the 2192nd position. Thus, the median is 0.

c) The median would still be 0, because there are still 2200 people who gave 0 as a response. The mean would now be 1.95.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2200(0) + 279(1) + 97(2) + 40(3) + 1767(4.5)}{4383} = \frac{8544.5}{4383} = 1.95$$

d) The median is the same for both because the median ignores much of the data. The data are discrete; hence, a high proportion of the data falls at only one or two values. The mean is better in this case because it uses the numerical values of all of the observations, not just the ordering.

2.45 Airplane crashes

In this case, the mean is likely to be more useful because it uses the numerical values of all of the observations, not just the ordering. Because so many airlines companies would report 0 airplane crashes, the median is not very useful. It ignores too much of the data.

Section 2.4 Measuring the Variability of Quantitative Data

2.46 Traffic Violations

a) The range is 8; this is the distance from the smallest to the largest observation. In this case, there are eight points separating the lowest and the largest number of points accumulated (8 – 0 = 8).

b) The standard deviation is the typical distance of an observation from the mean (which is 2.5).

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{69}{11} = 6.273$$

$$s = \sqrt{s^2} = \sqrt{6.273} = 2.505$$

The standard deviation of 2.505 indicates a typical number of points accumulated on their driving records are 2.505 points from the mean of 2.5.

2.46 (continued)

c) Redo (a) and (b).

a) The range is 20. This is the distance from the smallest to the largest observation. In this case, there are 20 points separating the lowest and largest number of points accumulated (20 – 0 = 20).

b) The standard deviation is the typical distance of an observation from the mean (which is 4).

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{348}{11} = 31.636$$

$$s = \sqrt{s^2} = \sqrt{31.636} = 5.625$$

The standard deviation of 5.625 indicates a typical number of points accumulated on their driving records is 5.625 points from the mean of 4. The range and mean both increase when an outlier is added.

2.47 Life expectancy

a) Upon examination of the data, the countries in Africa will have a larger standard deviation since the spread of the data is greater for this group than for the countries in Western Europe.

b) Western Europe:

$$\bar{x} = \frac{\sum x}{n} = \frac{81 + 80 + 80 + 81 + \cdots + 80 + 82 + 82 + 83}{15} = \frac{1220}{15} = 81.3333$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(81 - 81.3333)^2 + \cdots + (83 - 81.3333)^2}{15 - 1}} = 1.05$$

Africa:

$$\bar{x} = \frac{\sum x}{n} = \frac{47 + 50 + 51 + 57 + \cdots + 64 + 63 + 62 + 61}{16} = \frac{914}{16} = 57.125$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{(47 - 57.125)^2 + \ldots + (61 - 57.125)^2}{16 - 1}} = 5.18$$

Note that the standard deviation for the Western Europe group, 1.0 (rounded), is much smaller than for the Africa group, 5.2.

2.48 Life expectancy including Russia

We would expect the standard deviation to be larger since the value for Russia is significantly smaller than the rest of the group adding additional spread to the data. The standard deviation including Russia is, in fact, 3.01.

2.49 Shape of home prices?

The most plausible value is $60,000. –$15,000 is not possible because a standard deviation cannot be negative. $1,000 and $1,000,000 are unlikely because they are too small or too big, respectively, for a typical deviation. One would not expect the typical deviation to be that far from the median for home prices.

2.50 Exam standard deviation

The most realistic value is 12. There are problems with all the others.

–10: We can't have a negative standard deviation.

0: We know there is spread because the scores ranged from 35 to 98, so the standard deviation is not 0.

3: This standard deviation seems very small for this range.

63: This standard deviation is too large for a typical deviation. In fact, no score differed from the mean by this much.

2.51 Heights

a) According to the Empirical Rule, 68% of men would be within one standard deviation of the mean between 71 – 1(3) = 68 and 71 + 1(3) = 74 inches. 95% of men would be within two standard deviations of the mean, between 71 – 2(3) = 65 and 71 + 2(3) = 77 inches. All or nearly all men would be within three standard deviations of the mean, between 71 – 3(3) = 62 and 71 + 3(3) = 80 inches.

b) The mean for women is lower than the mean for men. Because each gender's heights would tend to be closer to that gender's mean than to the overall mean, the standard deviation would be smaller when we compared them with the appropriate gender group than when we compared them to the overall group. Would not expect unimodal but more bimodal.

2.52 Histograms and standard deviation

a) (i) The sample on the right has the largest standard deviation since is the most spread out.
(ii) The sample in the middle has the smallest standard deviation since it has no spread.

b) The Empirical Rule is relevant only for the distribution on the left because the distribution is bell-shaped.

2.53 On-time performance of airlines

68% of the 72 on time observed rates would be within one standard deviation from the mean: between 82.93% and 88.93%; 95% of the observations would be within two standard deviations from the mean: between 79.93% and 91.93%; all or nearly all time observed rates would be within three standard deviations from the mean: between 76.93% and 94.93%.

2.54 Students' shoe size

a) 95% of shoe sizes would fall within two standard deviations from the mean: between 9.91 – 2(2.07) = 5.77 and 9.91 + 2(2.07) = 14.05.

b) A student who is three standard deviations above the mean would have a shoe size of 9.91 + 3(2.07) = 16.12. This would be an unusual observation because typically all or nearly all observations fall within three standard deviations from the mean. In a bell-shaped distribution, this would likely be about the highest score one would obtain.

2.55 Shape of cigarettes taxes

With a bell-shaped distribution, we expect scores to extend about three standard deviations from the mean in either direction. The lowest possible value of 0, however, is only $\frac{0 - 73}{48} = -1.52$, or 1.52 standard deviations below the mean, and so the distribution likely is skewed to the right.

2.56 Empirical rule and skewed, highly discrete distribution

a)

$$\bar{x} = \frac{\sum x}{n} = \frac{8418(0) + 1594(1) + 10(2)}{10022} = \frac{1614}{10022} = 0.16$$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{8418(0 - 0.16)^2 + 1594(1 - 0.16)^2 + 10(2 - 0.16)^2}{10022 - 1}} = 0.37$$

2.56 (continued)

b)

	Observations	Predicted by Empirical Rule
One standard deviation from the mean is between 0.16 – 1(0.37) = –0.21 and 0.16 + 1(0.37) = 0.53	84.0%	68%
Two standard deviations from the mean is between 0.16 – 2(0.37) = –0.58 and 0.16 + 2(0.37) = 0.90	84.0%	95%
Three standard deviations from the mean is between 0.16 – 3(0.37) = –0.95 and 0.16 + 3(0.37) = 1.27	99.9%	About 100%

There are more observations within one standard deviation of the mean and fewer within two standard deviations than would be predicted by the Empirical Rule.

c) The Empirical Rule is only valid when used with data from a bell-shaped distribution. This is not a bell-shaped distribution; rather, it is highly skewed to the right. Most observations have a value of 0, and hardly any have the highest value of 2.

2.57 Time spent using electronic devices

These statistics suggest that this distribution is highly skewed toward the right for the two main reasons. The mean is larger than the median, and the standard deviation is almost one and half larger than the mean.

In fact, the lowest possible value of 0 is only $\frac{17}{11.5} = 1.48$ standard deviations below the mean.

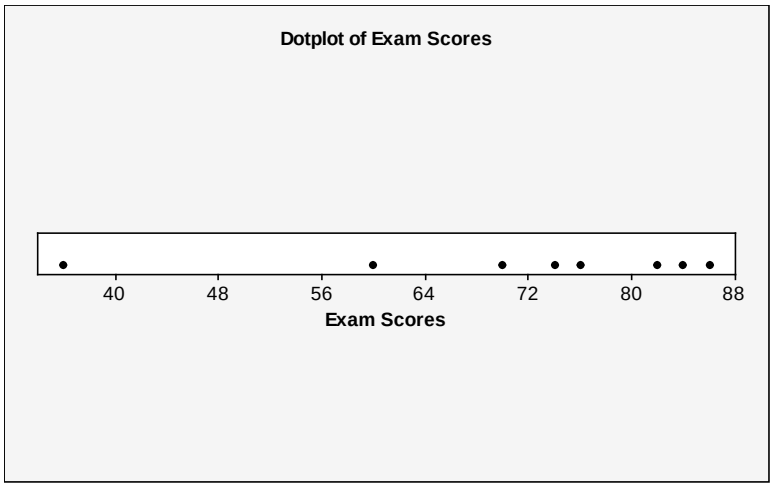
2.58 Facebook friends

a) The standard deviation is larger than one-half the mean; in addition, the mean is higher than the median. In fact, the lowest possible value of 0 is only $\frac{170}{90} = 1.89$ standard deviations below the mean. This is typical of a skewed right distribution with one or more potential outliers.

b) The Empirical Rules does not apply to these data because they do not appear to be bell-shaped.

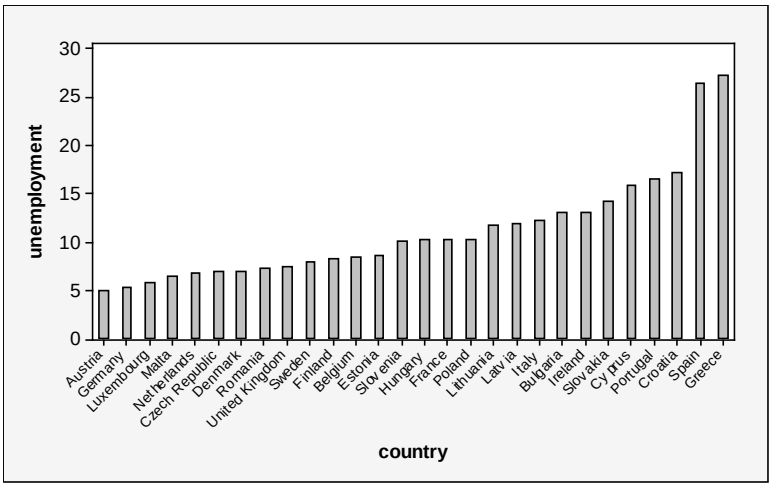
2.59 Judging skew using $\bar{x}$ and s

The largest observation is 86 is less than one standard deviation above the mean of 70.4. Specifically, 86 is only $\frac{86 - 70.4}{16.7} = 0.93$ standard deviations above the mean. The smallest observation is 35, which is $\frac{35 - 70.4}{16.7} = -2.12$, or 2.12 standard deviations below the mean. This distribution, therefore, is likely skewed to the left.



2.60 Youth unemployment in the EU

a)



b) Mean = 11.1%, median = 10.1%, s = 5.6%

c) The distribution of the youth unemployment rate in the EU is skewed to the right, with two countries (Greece and Spain) showing an unemployment rate of more than 20%. The mean unemployment rate is 11.1% (median 10.1%). The variability in the unemployment rate is relatively large, with a standard deviation of 5.6%, but this may be inflated due to the two outliers and right-skewness of the distribution.

2.61 Create data with a given standard deviation

a) One possible answer: 30, 50, 80

b) One possible answer: 10, 50, 90.

c) The largest standard deviations results from two 0s and two 100s, with s = 57.74.

Section 2.5 Using Measures of Position to Describe Variability

2.62 Vacation days

a) Median: Find the middle value of 13, 25, 26, 28, 34, 35, 37, 42.
The median is 31, the average of the two middle values, 28 and 34.

b) The first quartile is the median of 13, 25, 26, 28. The first quartile is 25.5, the average of the two middle values, 25 and 26.

c) The third quartile is the median of 34, 35, 37, 42. The third quartile is 36, the average of the two middle values.

d) 25% of countries have residents who take fewer than 25.5 vacation days, half of countries have residents who take fewer than 31 vacation days, and 75% of countries have residents who take fewer than 36 vacation days per year. The middle 50% of countries have residents who take an average of between 25.5 and 36 vacation days annually.

2.63 Youth unemployment

a) The median is 10.15, the average of the 14th and 15th values, 10.0 and 10.3. In 2013, half of the European Union nations had an unemployment rate less than 10.15%.

b) The first quartile is 7.15, the average of the 7th and 8th values, 7.0 and 7.3. In 2013, 75% of the European Union nations had an unemployment rate larger than 7.15% (or 25% had rate less than 7.15%).

c) The third quartile is 13.05, the average of the 21st and 22nd values, 13.0 and 13.1. In 2013, the unemployment rate was larger than 13.05% for 25% of the European Union nations.

d) The 10th percentile will be around 6% because Q1 = 7.15%.

2.64 On-time performance of airlines

a) One-fourth had an on time arrival rate less than 83.75, and one-fourth had an on time arrival rate greater than 87.75.

b) The mean and median are the about the same, and the first and third quartiles are equidistant from the median. These are both indicators of a roughly symmetric distribution.

2.65 Students' shoe size

a) One quarter had shoe size below 8 and one quarter had shoe size above 11.

b) The mean and median are about the same, but the first quartile is slightly more from the median than the third quartile. The latter is an indicator of a roughly skewed left distribution.

2.66 Ways to measure variability

a) The range is even more affected by an outlier than is the standard deviation. The standard deviation takes into account the values of all observations and not just the most extreme two.

b) With a very extreme outlier, the standard deviation will be affected both because the mean will be affected and because the deviation of the outlier (and its square) will be very large. The IQR would not be affected by such an outlier.

c) The standard deviation takes into account the values of *all* observations and not just the two marking 25% and 75% of observations.

2.67 Variability of net worth of billionaires

a) (i) Q_1, marking the lowest 25% of states, has a value of 3.6 billion. Thus, 75% of billionaires have wealth greater than 3.6 billion.

(ii) Q_3, marking the highest 25% of states, has a value of 8.2 billion. Thus, 25% of billionaires have wealth greater than 8.2 billion.

b) The two values that demarcate the middle 50% are $Q_1 = 3.6$ billion and $Q_3 = 8.2$ billion.

c) The interquartile range (IQR) is the difference between Q_1 and Q_3. IQR = $Q_3 - Q_1 = 8.2 - 3.6 = 4.6$ billion. For the middle 50% of billionaires' wealth, $4.6 billion is the distance between the largest and smallest wealth amount.

d) With a bell-shaped distribution, we expect Q_1 and Q_3 to be roughly equidistant from the median which is not the case here. The maximum value is also quite far from Q_3. Thus, it appears that the distribution is skewed to the right.

2.68 Traffic violations

a) The range is 8. This is the distance from the smallest to the largest observation. In this case, there are 8 points separating the smallest and the largest number of points accumulated (8 – 0 = 8).

b) The interquartile range is the difference between Q_3 and Q_1. IQR = $Q_3 - Q_1 = 4$.

c) Redo (a) and (b).

a) The range is 20. This is the distance from the smallest to the largest observation. In this case, there are 20 points separating the smallest and the largest number of points accumulated (20 – 0 = 20).

b) Q_1, the median of all scores below the median, is still 0. Q_3, the median of all scores above the median, becomes 4.5. The interquartile range becomes the IQR = $Q_3 - Q_1 = 4.5$.

The IQR is least affected by the outlier because it does not take the magnitudes of the two extreme scores into account at all, whereas the ranges do.

2.69 Infant mortality Africa

a) Q1 is the median of the lower half of the sorted data: 54, 63, 68, 76, 78, 79, 80. It is 76. Q3 is the median of the upper half of the sorted data: 81, 84, 96, 101, 110, 121, 154. It is 101.

b) IQR = Q3 – Q1 = 25. For the middle half of the infant mortality rates, the distance between the largest and smallest rates is 25.

2.70 Infant mortality Europe

Q1 is the median of the lower half of the sorted data: 3, 3, 3, 4, 4, 4, 4. It is 4. Q2, the median, is the average of the middle two data values $\frac{4+4}{2} = 4$. Q3 is the median of the upper half of the sorted data: 4, 4, 4, 4, 5, 5, 5. It is 4.

2.71 Computer use

a) This five-number summary suggests that the distribution is skewed to the right. The distance between the minimum and the median is much smaller than the distance between the median and the maximum.

b) In this case, outliers would be those values more than 1273.5 points from the first and third quartiles:
IQR = Q3 – Q1 = 1105 – 256 = 849 and 1.5(IQR) = 1.5(849) = 1273.5
The lower boundary: Q1 – 1.5(IQR) = 256 – 1273.5 = –1017.5
The upper boundary: Q3 + 1.5(IQR) = 1105 + 1273.5 = 2378.5
In the current example, the lowest score is 4, so there are no scores below –1017.5. The highest score, on the other hand, is 320,000, much higher than 2378.5. Thus, there are potential outliers according to this criterion.

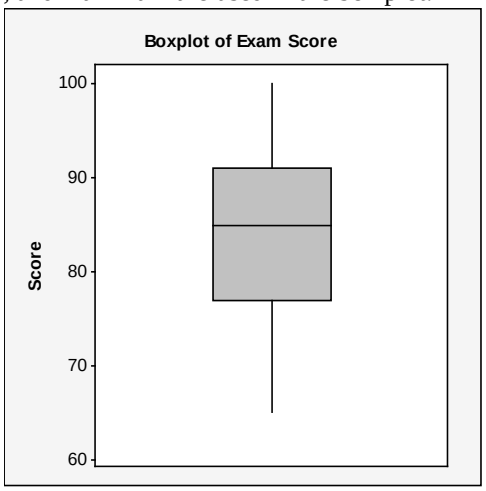
2.72 Central Park temperature distribution revisited

a) We would expect it to be skewed to the left because the maximum is closer to the median than is the minimum.

b) Numbers are approximate: Minimum: 49.0, Q1: 52.5, Median: 53. 5, Q3: 55.0, Maximum: 57.0
These approximations support the premise that the distribution is skewed to the left if it is skewed. The median is closer to the maximum and Q3 than it is to the minimum and Q1.

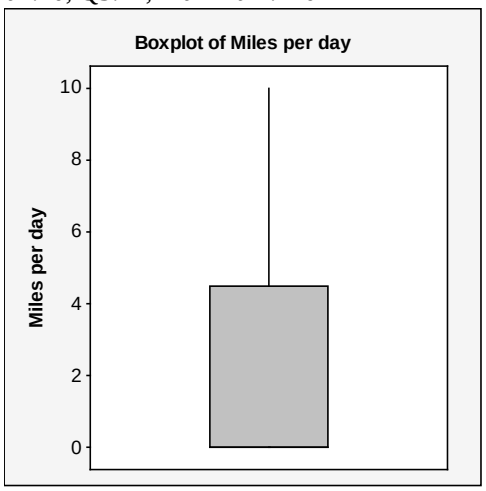
2.73 Box plot for exam

The minimum, Q1, median, Q3, and maximum are used in the box plot.



2.74 Public transportation

a) Minimum: 0, Q1: 0, Median: 0, Q3: 4, Maximum: 10



2.74 (continued)

b) Q1 and the median share the same line in the boxplot because so many employees have a score of zero that the middle score of the whole set of data is zero and the middle score of the lower half of the data also is zero.

c) There is no whisker because the minimum score also is zero. This situation resulted because there are so many people with the lowest score.

2.75 Energy statistics

a) Numbers are approximate: Minimum: 50, Q1: 130, Median: 160, Q3: 250, Maximum: 650
One country was a potential outlier, the one around 650.

b) We can know how far Italy was from the mean in terms of standard deviations by calculating its *z*-score. It is 0.47 standard deviations below the mean of 195.

$$z = \frac{x - \bar{x}}{s} = \frac{139 - 195}{120} = -0.47$$

c) The U.S. is 1.16 standard deviations above the mean.

$$z = \frac{x - \bar{x}}{s} = \frac{334 - 195}{120} = 1.16$$

2.76 European Union youth unemployment rates

a) In a box plot, Q1 = 7.15 and Q3 = 13.05, would be the outer edges of the box. 1.5(IQR) = 1.5(13.05 – 7.15) = 8.85. The whisker on the left would extend to the minimum 4.9, since it is larger than 7.15 – 8.85 = –1.7. The whisker on the right would extend to the value of 17.2 (Croatia), since it is the largest value below 7.15 + 8.85 = 21.9.

b) Greece (27.3) and Spain (26.4) have values larger than 21.9, so would be considered outliers.

c) Greece's score is 2.89 standard deviations above the mean, and thus, is not an outlier according to the three standard deviation criterion.

$$z = \frac{x - \bar{x}}{s} = \frac{27.3 - 11.1}{5.6} = 2.89$$

d) A *z*-score of 0 indicates that the country's unemployment rate is zero standard deviations from the mean; hence, the unemployment rate is equal to the mean. In this case, a country with an unemployment rate of 11.1 would have a z-score of 0.

2.77 Air pollution

a) Finland's pollution is exactly one standard deviation above the mean pollution of all countries in the EU.

$$z = \frac{x - \bar{x}}{s} = \frac{11.5 - 7.9}{3.6} = 1$$

b) Sweden's pollution is 0.64 standard deviation below the mean pollution of all countries in the EU.

$$z = \frac{x - \bar{x}}{s} = \frac{5.6 - 7.9}{3.6} = -0.64$$

c) The United Kingdom's pollution is exactly equal to the mean pollution of all countries in the EU.

$$z = \frac{x - \bar{x}}{s} = \frac{7.9 - 7.9}{3.6} = 0$$

2.78 Height of buildings

a) $$z = \frac{x - \bar{x}}{s} = \frac{761 - 195.57}{106.32} = 5.32$$

b) The positive sign indicates that the height of 761 feet is above the mean.

c) Because the height of 761 feet is more than three standard deviations from the mean, it is a potential outlier.

2.79 Marathon results

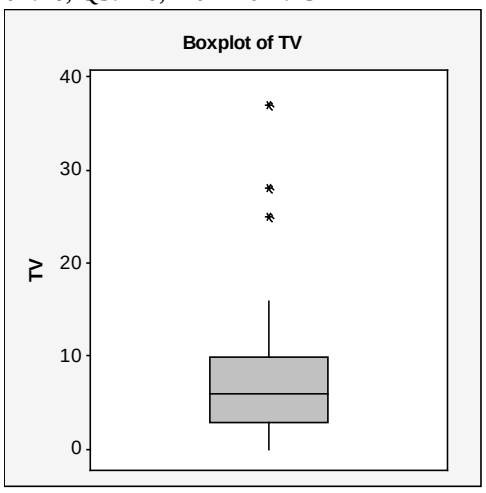
The appropriate statistic here would be the *z*-score.

$$z = \frac{x - \bar{x}}{s} = \frac{9377 - 9073}{161.4} = 1.88.$$

This *z*-score indicates that the finish time of the last finalist is only 1.88 standard deviations from the mean, which is less than three standard deviations above the mean, and thus, it would not be a potential outlier – in other words, an usual result.

2.80 Florida students again

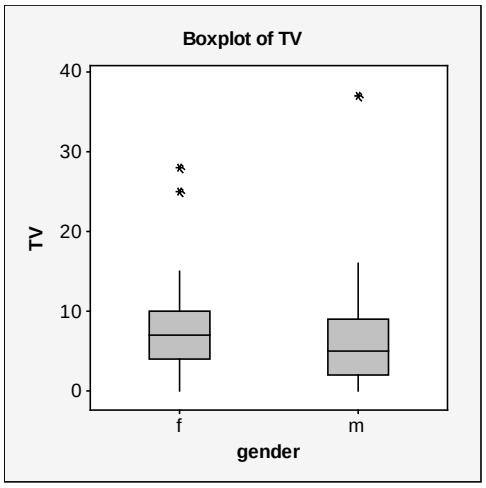
a) The distribution depicted in the box plot is skewed to the right. Most observations fall between about 0 and 15 but there are a few outliers representing very large values.
Minimum: 0, Q1: 3, Median: 6, Q3: 10, Maximum: 37



b) Since IQR = 10 – 3 = 7 and 1.5(IQR) = 1.5(7) = 10.5, the 1.5(IQR) criterion would indicate that all data should fall between about 3 – 10.5 = –7.5 and 10 + 10.5 = 20.5. Because some data points fall beyond this range, it appears that there are potential outliers.

2.81 Females or males watch more TV?

Based on the Florida survey data, females tend to watch more TV. The median, Q1 and Q3 are higher for females than for males.



2.82 CO_2 comparison

a) The two outliers for Central and South America have a value of roughly 12 metric tons.

b) The distributions would be skewed to the right. The median sits low in the box (pulled toward the first quartile), and the upper whisker stretches out more from the box to the maximum value than the left whisker stretches to minimum value.

c) The median CO_2 emission is much larger in Europe than the median for Central and South America. The spread of the middle 50% of the distribution of emissions, as measured by the IQR, seems to be about the same for Europe and Central and South America. 75% of the distribution of emissions for Europe is higher than the lower 75% of the distribution of emissions for Central and South America. Overall, emissions are higher for Europe than for Central and South America.

Section 2.6 Recognizing and Avoiding Misuses of Graphical Summaries

2.83 Cell phone bill

a) The mean is $105.43 and the median is $92.

b) It is misleading because the mean is so heavily influenced by the outlier (the highest cell phone bill) which it is not a typical value. The median would be a much more accurate summary of these bills.

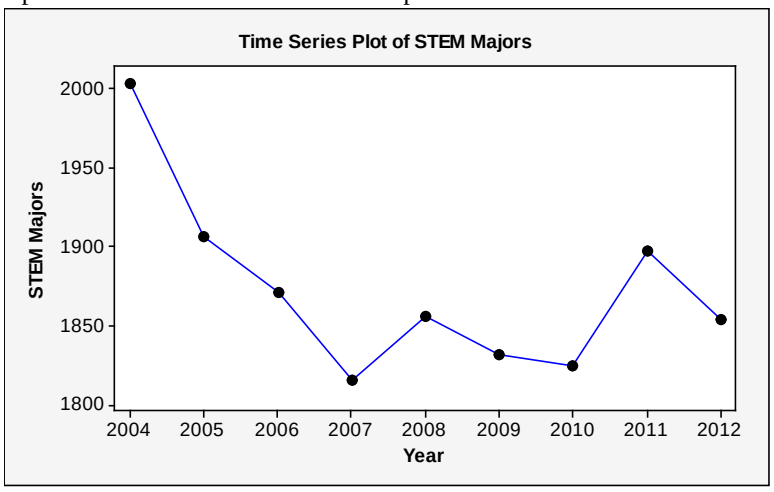
2.84 Market share for food sales

a) One problem with this chart is that the percentages do not add up to 100. Second, the Tesco slice seems too large for 27.2%. A third problem is that contiguous colors are very similar. This increases the difficulty in easily reading this chart.

b) It would be easier to identify the mode with a bar graph because one would merely have to identify the highest bar.

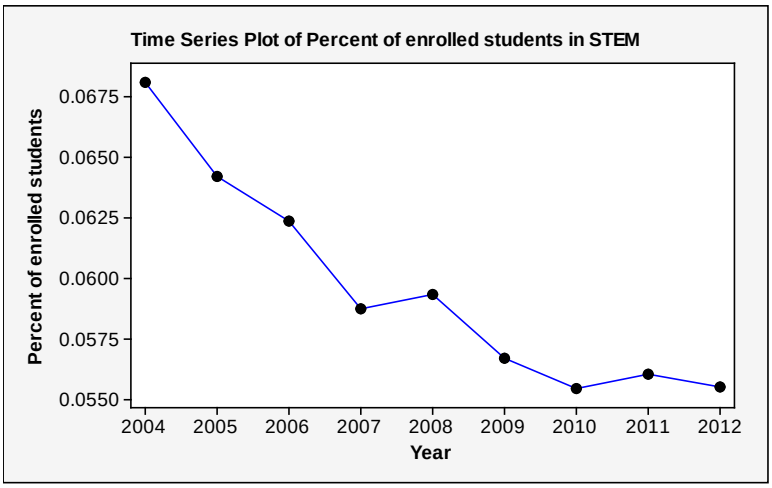
2.85 Enrollment trends

a) This graph shows an overall decrease in enrollment in STEM majors at first, with what appears to be somewhat of a plateau toward the end of the time span.



2.85 (continued)

b) This graph shows a gradual decrease over time in the percentage of students who are enrolled in STEM majors.



c) The graphs in (a) and (b) tell us that although there are some fluctuations in the numbers of students enrolling in STEM majors over the years, there is a steady decrease in the percentage of enrolling students who are enrolled in STEM majors over the years. We cannot learn this from Figures 2.18 and 2.19.

2.86 Terrorism and war in Iraq

a) This graph is misleading. Because the vertical axis does not start at 0, it appears that six times as many people are in the "no, not" column than in the "yes, safer" column, when really it's not even twice as many.

b) With a pie chart, the area of each slice represents the percentage who fall in that category. Therefore the relative sizes of the slices will always represent the relative percentages in each category.

2.87 BBC license fee

The 2013 projection is shown where the observation would be plotted for the year 2007, not 2013.

2.88 Federal government spending

The slices do not seem to have the correct sizes, for instance the slice with 16% seems larger than the slice with 19%.

2.89 Bad graph

Answers will vary.

Chapter Problems: Practicing the Basics

2.90 Categorical or quantitative?

a) Number of children in family: quantitative

b) Amount of time in football game before first points scored: quantitative

c) Choice of major (English, history, chemistry, …): categorical

d) Preference for type of music (rock, jazz, classical, folk, other): categorical

2.91 Continuous or discrete?

a) Age of mother: continuous

b) Number of children in a family: discrete

c) Cooking time for preparing dinner: continuous

d) Latitude and longitude of a city: continuous

e) Population size of a city: discrete

2.92 Young non-citizens in the U.S.

a) Region of Birth is Categorical.

Noncitizens aged 18 to 24 in the United States

Region of Birth	Number (in Thousands)	Percentage
Africa	115	$\frac{115}{2568} = 0.045 = 4.5\%$
Asia	590	$\frac{590}{2568} = 0.230 = 23.0\%$
Europe	148	$\frac{148}{2568} = 0.058 = 5.8\%$
Latin America & Caribbean	1666	$\frac{1116}{2568} = 0.649 = 64.9\%$
Other	49	$\frac{49}{2568} = 0.019 = 1.9\%$
Total	**2568**	

b) Mode. Most young noncitizens are from Latin America and the Caribbean.

c) Latin America and Caribbean, Asia, Europe, Africa, Other. One immediately sees that most young noncitizens are from two regions, Latin America and Caribbean and from Asia.

2.93 Cool in China

a) The variable being measured is the personality trait that defines "cool."

b) This is a categorical variable.

c) Because the data are categorical with unordered categories, we could use only the bar chart and the modal category.

2.94 Chad voting problems

a) We first locate the dot directly above 11.6% on the horizontal *x* axis. We then look at the vertical *y* axis across from this point to determine the label for that dot: Optical scanning with a two-column ballot. This tells us that the over-vote was highest among those using optical scanning with a two-column ballot.

b) We first locate the dots above the lowest percentages on the *x* axis. We then determine the labels across from these dots on the *y* axis to determine the lowest two combinations: optical, one column, and votomatic, one column. Thus, the lowest over-voting occurred when voters had a ballot with only one column that was registered either using optical scanning or votomatic (manual punching of chads).

c) We could summarize these data further by using a bar for each combination: optical, one column; optical, two column; votomatic, one column, etc. For each bar, we could then plot the average over-vote of all counties in that category. To do this, we would need the exact percentages of each county in each category.

2.95 Number of siblings

a) The sample size is the sum of all counts, and in this case, it is 3003.

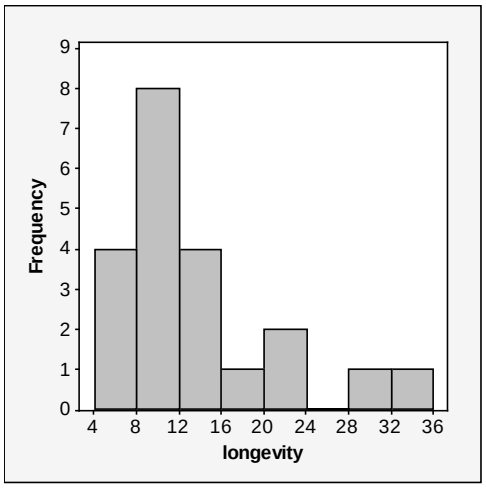
b) The most appropriate graph would be a histogram because there are so many data points. Dot plots and stem-and-leaf plots are unwieldy with many data points.

c) A histogram would show that data are skewed to the left. There are many data points at higher numbers, but the number of observations decreases rapidly as we get to lower numbers of siblings.

2.96 Longevity

a)
0 |
0 | 57789
1 | 0011112234
1 | 9
2 | 23
2 |
3 | 0
3 | 5

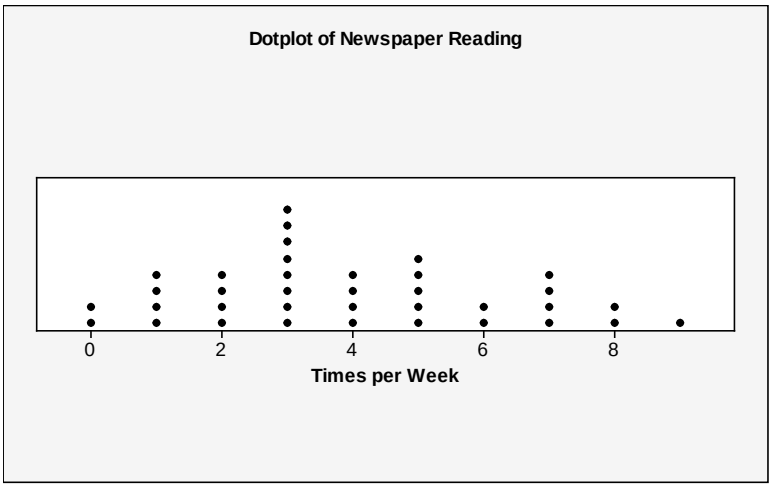
b)



c) The distribution of longevity is right-skewed. Most animals live to be between 5 and 15 years old.

2.97 Newspaper reading

a)



2.97 (continued)

b)
0 | 00
1 | 0000
2 | 0000
3 | 00000000
4 | 0000
5 | 00000
6 | 00
7 | 0000
8 | 00
9 | 0

The leaf unit is identified above. The stems are the whole numbers, 0 through 9.

c) The median is the middle number. There are 36 numbers, so the median is between the 18th and 19th which have the values 3 and 4, respectively. Thus, the median is 3.5.

d) The distribution is slightly skewed to the right.

2.98 Match the histogram

a) symmetric and bimodal
b) skewed to the right
c) skewed to the left
d) symmetric and unimodal

2.99 Sandwiches and protein

a)
0 | 8
1 |
1 | 7889
2 | 113
2 | 666

b) A stem-and-leaf plot allows one to see the individual amounts.

c) The protein amounts are mostly between 17 and 21 grams with a few sandwiches having a higher protein value of 26 grams. There appears to be one outlier having only 8 grams of protein.

2.100 Sandwiches and cost

a) (The data values have been truncated.)
2 | 4
2 | 999
3 | 1444
3 | 688

b) A stem-and-leaf plot allows one to see the individual prices.

c) Most of the sandwiches cost between \$2.90 and \$3.89. The prices are skewed to the left with one sandwich costing only \$2.49.

2.101 What shape do you expect?

a) Number of times arrested in past year – skewed to the right because most values are at 0 but there are some large values.

b) Time needed to complete difficult exam (maximum time is 1 hour) – skewed to the left because most values are at 1 hour or slightly less, but some could be quite a bit less.

c) Assessed value of home – skewed to the right because there are some extremely large values.

d) Age at death – skewed to the left because most values are high, but some very young people die.

2.102 Sketch plots

NOTE: Plots will vary, but should have the following characteristics.

a) It would be skewed to the right, and the mean would be greater than the median because of a few mansions that sell for millions.

b) It would be skewed to the right, and the mean would be higher than the median. Most women do not give birth over age 40. Thus, the median would be zero. The mean, however, would be positive, because some women do give birth over the age of 40.

c) It would be skewed to the left, and the mean would be lower than the median. The mean would be pulled down by the outlier of 50. The standard deviation is only 10, so there probably aren't lots of low scores. Moreover, the highest possible score of 100 is only 12/10 = 1.2 standard deviations above the mean.

d) It would be skewed to the left, and the mean would likely be lower than the median. Most people with cars drive them every month, but a few drive them less, and some hardly or not at all. These outliers would pull the mean, but not the median, lower. The median and mode probably would be 12.

2.103 Median versus mean sales price of new homes

We would expect the mean sales price to have been higher due to the distribution being skewed to the right. A few very expensive homes will greatly affect the mean, but not the median sales price.

2.104 Household net worth

a) The distribution of these families' net worth is likely to be skewed to the right because relatively few families would have very high net worth so that we expect the mean to be greater than the median.

b) When assets such as homes and retirement savings decline due to a recession, it is typical for the highest valued assets to be affected the most. Thus, we would expect the mean net worth to drop more than the median net worth.

2.105 Golfers' gains

a) The data for the 90 players would be skewed to the right with the majority of the golfers earning between \$1 and \$3 million and a few earning over \$3 million.

b) Since the data is skewed to the right, the mean would be the higher value of \$2,090,012 and the median the lesser value of \$1,646,853.

2.106 Hiking

The classification into easy, medium or hard is categorical and the length classification is quantitative.

2.107 Lengths of hikes

a) One example is 1, 2, 4, 6, 7. Both the mean and median are 4.

b) One example is 2, 2, 3, 5 and 6.

2.108 Central Park monthly temperatures

a) Both distributions are fairly symmetric and bell-shaped, with January having greater variability than July.

b) The mean temperature for January is around 32º and the mean temperature for July is around 76º. The average monthly temperature in January is approximately 44º less than the average monthly temperature in July.

c) The average monthly temperature in January is more variable than in July. The range of average temperatures for January is approximately 22º to 43º and the standard deviation is approximately 5º. The range of average temperatures for July is approximately 71º to 81º and the standard deviation is approximately 2º. It may be a bit surprising to see how much more variable are the average monthly temperatures in January than in July.

2.109 What does *s* equal?

a) Given the mean and range, the most realistic value is 12. –10 is not realistic because standard deviation must be 0 or positive. Given that there is a large range, it is not realistic that there would be almost no spread; hence, the standard deviation of 1 is unrealistic. 60 is unrealistically large; the whole range is hardly any more than 60.

b) –20 is impossible because standard deviations must be nonnegative.

2.110 Female heights

a) According to the Empirical Rule, 95% of scores in a bell-shaped distribution fall within two standard deviations of the mean.

$$\bar{x} - 2s = 65 - 2(3.5) = 58$$

$$\bar{x} + 2s = 65 + 2(3.5) = 72$$

Thus, 95% of heights likely fall between 58 and 72 inches.

b) The height for a woman who is three standard deviations below the mean is 54.5.

$$\bar{x} - 3s = 65 - 3(3.5) = 54.5$$

This is on the cusp of what would be considered an outlier according to the *z*-score criterion. Scores that are beyond three standard deviations from the mean are considered to be potential outliers. So, yes, this height is bordering on unusual.

2.111 Energy and water consumption

a) The distribution is likely skewed to the right because the maximum is much farther from the mean than the minimum is, and also because the standard lowest possible value of 0 is only 780/506 = 1.54 standard deviations below the mean.

b) The distribution is likely skewed to the right because the standard deviation is almost as large as the mean, and the smallest possible value is zero, only 1.15 standard deviation below the mean.

2.112 Hurricane damage

a) The distribution is skewed to the right.

b) The median should be used since the distribution is skewed to the right.

c) The values are correct.

d) The distribution of hurricane damage is skewed to the right, with the damage for the costliest hurricane Katrina (more than 100 billion) far exceeding all others. The median damage was 7.9 billion. The 25% most costly hurricanes had a cost of over 11.8 billion, whereas the 25% fewest damaging hurricanes cost no more than 5.7 billion.

2.113 More hurricane damage

a) The percentage differs from 68% because of the extreme right skew of the distribution.

b) The mean and standard deviation would get much smaller due to the removal of the extreme value. The median, IQR, and 10th percentile would not change much.

2.114 Income statistics

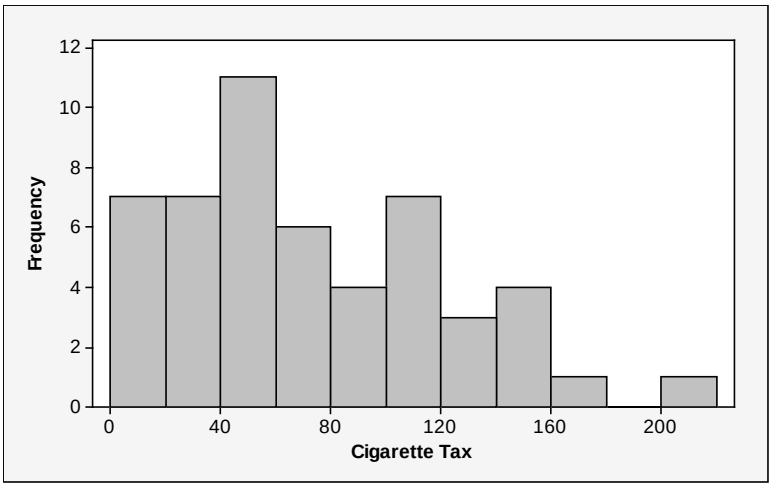
a) For a right-skewed distribution, such as the income of all people, the Empirical Rule is not appropriate (the Empirical Rule states that all or nearly all scores will fall within three standard deviations of the mean ($\bar{x} \pm 3s$). In this case, we could expect many scores above $222,900.

b) The center for a female is about $19,370 less than the center for a male. The variability for males is larger. Up to some proportionality, these distributions are likely similar in shape, they are centered around different values.

c) The *z*-score for a male is $z = \dfrac{x - \bar{x}}{s} = \dfrac{55000 - 47836.1}{58353.55} = 0.12$ and the *z*-score for a female is $z = \dfrac{x - \bar{x}}{s} = \dfrac{45000 - 28466}{36961.1} = 0.45$. Thus, a female's income is relatively higher than a male's income.

2.115 Cigarette tax

a)



The histogram shows a unimodal distribution that is skewed to the right. If there are any outliers, they would be the most extreme scores, such as the one around 200.

b) The mean is 72.85 and the median is 60. The mean is inflated relative to the median as one would expect from the distribution depicted in the histogram that is skewed to the right. The few high scores would pull the mean higher, but not the median.

c) The standard deviation is 48.00. This indicates that the typical score falls about 48.0 from the mean.

2.116 Cereal sugar values

a) Numbers are approximate: Minimum: 0, Q1: 4, Median: 9.5, Q3: 13.5, Maximum: 18

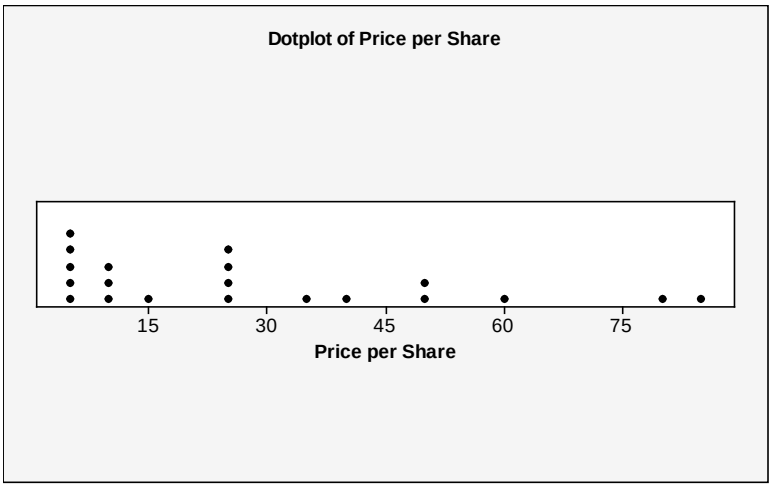
b) Because the median is closer to Q3 and the maximum than it is to Q1 or the minimum, it appears that this distribution is slightly skewed to the left.

c) This sugar value falls 1.64 standard deviations below the mean.

$$z = \frac{x - \bar{x}}{s}, \quad z = \frac{0 - 8.75}{5.32} = -1.64$$

2.117 NASDAQ stock prices

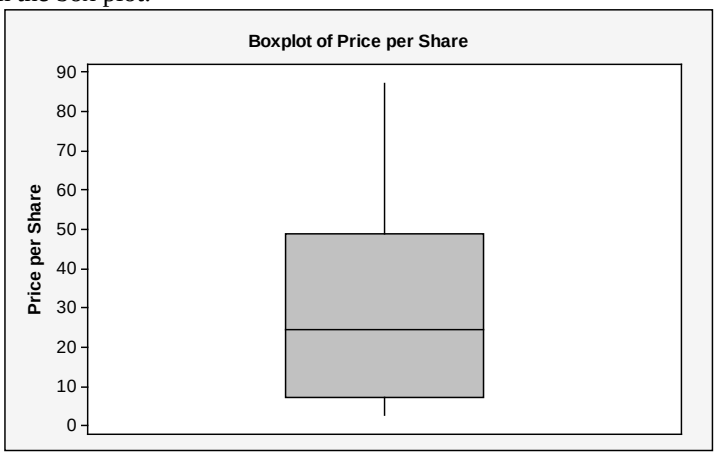
a)



b) The median is the average of the two middle numbers, 23 and 26. Thus, the median is 24.5. The first quartile is the median of all the numbers below the median: 3, 4, 4, 7, 7, 8, 9, 9, 13, and 23. Thus, the first quartile is 7.5. The third quartile is the median of all numbers above the median: 26, 26, 26, 37, 40, 52, 52, 60, 78, 87. Thus, the third quartile is 46.

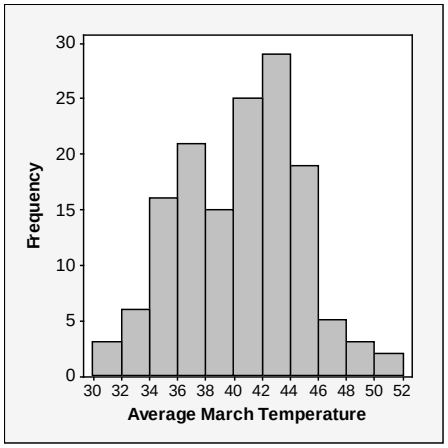
2.117 (continued)

c) The box plot does not show the gaps in the observations. Also, the individual data values cannot be reproduced from the box plot.



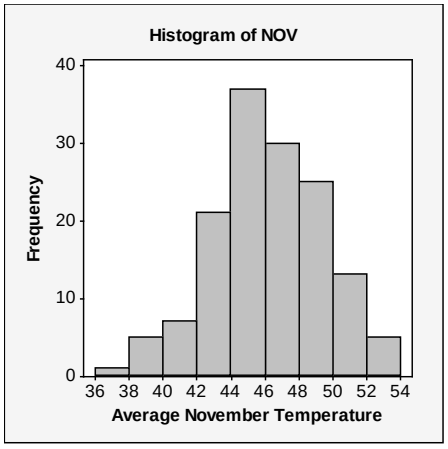
2.118 Temperatures in Central Park

a) The distribution appears to be fairly symmetric, but perhaps a little left-skewed . Most values lie between 35 and 45 degrees, but range from 30 to 51 degrees. The mean and median are almost the same; the typical March temperature is around 42 degrees. The spread is not large in comparison to the mean.



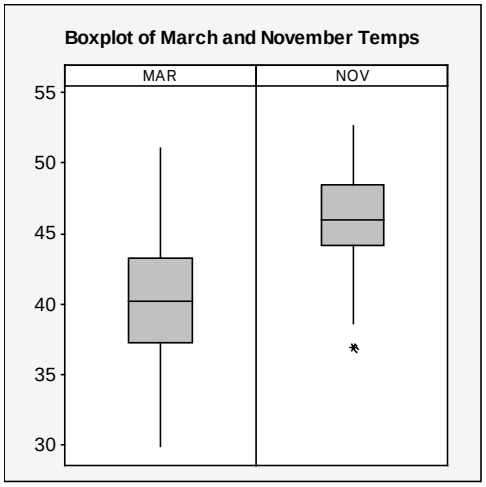
b) Mean: 40.4; Standard deviation: 4.2

c) The mean average temperature is higher in November than in March and the spread of average temperatures is less in November than in March.



2.118 (continued)

d) As indicated by the histograms, the average monthly temperature is greater in November than in March and the standard deviation is less. The side-by-side box plot makes it easier to see the relative spreads of the data as well as the difference in the means.



2.119 Teachers' salaries

a) The range is the maximum minus the minimum. Range = 69,119 – 35,070 = 34,049
The interquartile range (IQR) equals Q3 – Q1. IQR = 55,820 – 45,840 = 9980
These statistics indicate that the salaries range across a $34,049 span, and that the middle 50% of salaries range across an $9,980 span.

b) (i) The values at the ends of the boxes would be 45,480 (Q1) and 55,820 (Q3).
(ii) The line in the middle of the box would be the median, 48,630.
(iii) The lower end of the left whisker would be the minimum, 35,070.
(iv) The upper end of the right whisker would be the maximum, 69,119.

c) The minimum and Q1 are closer to the median than are Q3 and the maximum. This indicates that the data are likely skewed to the right.

d) The most realistic standard deviation would be 7000. 100 and 1,000 are too small for typical deviations given a range of 34,049. 25,000 is too big given that it is almost three-quarters of the range; the typical score could never be this far from the mean. 7000 is the only realistic score.

2.120 Health insurance

a) The distribution is most probably skewed to the right because the distance of Q3 from the median and from Q3 to the maximum is longer than the distance of Q1 from the median and Q1 to the minimum.

b) The most plausible value for the standard deviation of this distribution is 1000. The middle 50% of scores fall within a range of 681, making it plausible that the typical score would deviate about 1000 from the mean. We cannot have a negative percentage point, so –160 is not plausible. We know that there is variation, so 0 is not plausible. The whole range is not much more than 5016; thus, 5000 is implausibly large for the standard deviation of this distribution.

2.121 What box plot do you expect?

Box plots will vary, but should have the following characteristics.

a) The center of these data is closer to the maximum than the minimum. Although the mean is likely to be pulled by an outlier more than the median, this still indicates that the data might be skewed to the left, and that the box plot might have more distance between the median and both Q1 and the minimum than between the median and both Q3 and the maximum.

b) IQ scores are designed to be symmetric, and these data support that. The box plot, thus, would appear symmetric.

c) The mean is higher than is the median, indicating that the data are skewed to the right. Thus, the box plot would have more distance between the median and both Q3 and the maximum than between the median and both Q1 and the minimum.

2.122 High school graduation rates

a) The range is the difference between the lowest and highest scores: 91.8 – 79.9 = 11.9. The interquartile range (IQR) is the difference between scores at the 25th and 75th percentiles: IQR = Q3 – Q1 = 89.8 – 84.0 = 5.8

b) Potential outliers are 1.5(IQR) = 1.5(5.8) = 8.7 below Q1 or above Q3. This criterion suggests that potential outliers would be those scores less than 84.0 – 8.7 = 75.3 and greater than 89.8 + 8.7 = 98.5. There are no scores beyond these values, and so it would not indicate any potential outliers.

c) No, since the *z*-scores of both the minimum and maximum values are both less than 3 in absolute value. No outliers are present.

$$z = \frac{x - \bar{x}}{s} = \frac{79.9 - 86.9}{3.4} = -2.06 \qquad z = \frac{x - \bar{x}}{s} = \frac{91.8 - 86.9}{3.4} = 1.44$$

2.123 SAT scores

a) Because the right whisker extends further than does the left whisker, and the line through the center of the box is left of center, the box plot suggests that the distribution is somewhat skewed to the right.

b) Numbers are approximate: Minimum: 1400, Q1: 1475, Median: 1550, Q3: 1700, Maximum: 1800 The lowest score is approximately 1400 and the highest is 1800. The score below which the lowest 25% fall is approximately 1475, and the score above which the highest 25% fall is approximately 1700. The middle score, that below which 50% of the scores fall, is 1550.

c) If only viewing the box plot, we would not see that the distribution may be bimodal.

2.124 Blood pressure

a) A *z*-score of 1.19 indicates that a person with a blood pressure of 140, the cutoff for having high blood pressure, falls 1.19 standard deviations above the mean.

$$z = \frac{x - \bar{x}}{s} = \frac{140 - 121}{16} = 1.19$$

b) About 95% of all values in a bell-shaped distribution fall within two standard deviations of the mean – in this case, 32. About 95% of systolic blood pressures fall between 121 – 2(16) = 89 and 121 + 2(16) = 153.

2.125 Price of diamonds

The price of the diamond sold at $25 million falls –0.59 standard deviations below the mean.

$$z = \frac{x - \bar{x}}{s} = \frac{25 - 114.36}{150.28} = -0.59.$$

2.126 Who was Roger Maris?

a) Neither the minimum nor the maximum score reaches the criterion for a potential outlier of being more than three standard deviations from the mean (i.e., having a *z*-score less than –3 or greater than 3). Thus, there are no potential outliers according to three standard deviation criterion.

$$z = \frac{x - \bar{x}}{s} = \frac{5 - 22.92}{15.98} = -1.12 \qquad z = \frac{x - \bar{x}}{s} = \frac{61 - 22.92}{15.98} = 2.38$$

b) The maximum is much farther from the mean and median than is the minimum, an indicator that the distribution might not be bell-shaped. Moreover, the lowest possible value of 0 is only 22.92/15.98 = 1.43 standard deviations below the mean.

c) Based on the criteria noted above, this is not unusual. It does not even come close to meeting the three standard deviation criterion for a potential outlier and therefore is not an unusual number of homeruns for Roger Maris.

$$z = \frac{x - \bar{x}}{s} = \frac{13 - 22.92}{15.98} = -0.62$$

Chapter Problems: Concepts and Investigations

2.127 Baseball's great homerun hitters

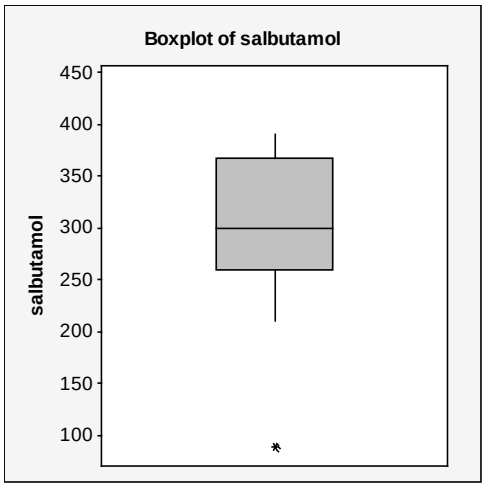
The responses will be different for each student depending on the methods used.

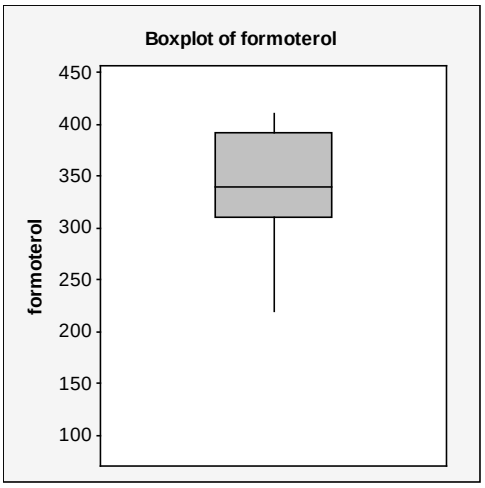
2.128 How much spent on haircuts?

The responses will be different for each student depending on the methods used.

2.129 Controlling asthma

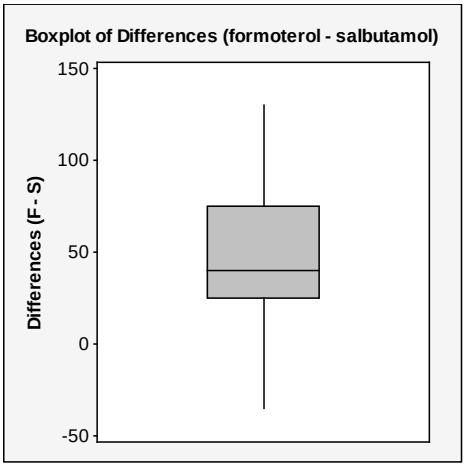
a) The distribution of children on both Formoterol (F) and Salbutamol (S) are skewed to the left. There is a data point that qualifies as an outlier, indicated by a dot to the far left, in the Salbutamol distribution. Children on Formoterol seem to be doing better, on average, than do children on Salbutamol.





b) Here are the difference scores for each child. A positive difference indicates a higher score for Formoterol than for Salbutamol.

Child	Formoterol	Salbutamol	Difference
1	310	270	40
2	385	370	15
3	400	310	90
4	310	260	50
5	410	380	30
6	370	300	70
7	410	390	20
8	320	290	30
9	330	365	–35
10	250	210	40
11	380	350	30
12	340	260	80
13	220	90	130



2.129 (continued)

If there is, on average, no difference between PEF levels for the two brands, the distribution of differences would be centered around 0, a score indicating no difference. The current difference scores appear skewed slightly to the right. The difference scores indicate a positive difference, on average. The center is well into the positive side, and the data points are quite spread out. Thus, children on Formoterol have higher scores, on average, than when they are on Salbutamol, although there is a quite a bit of variation in amount of improvement. Moreover, there appears to be one outlier, a child who responds more poorly on F than on S.

2.130 Google trend

The response to this exercise will be different for each student.

2.131 Youth unemployment by gender

The median unemployment rate is similar (at about 10%) for males and females. Also, for both genders, unemployment rate varies among countries (except Greece and Spain) from roughly 6% to 16%. The middle 50% of the distribution for females (IQR about 5%) has less variability than the distribution for males (IQR about 7%). For the two countries with highest unemployment rate (Greece and Spain, with rate larger than 20%), unemployment rate for females is even higher than for males. Including the outliers, both the male and female distributions are right skewed.

2.132 You give examples

Answers will vary.

a) Approximately symmetric – number of letters that can be remembered in a memory task, or IQ.
b) Skewed to the right – number of alcoholic beverages consumed in a week (this would be skewed by a few extreme binge-drinkers) or distance traveled to work (skewed by a few with incredibly long commutes).
c) Skewed to the left – happiness levels on one's wedding day (most would be very happy, but there'd likely be a few who were sad) or score on an easy exam (skewed by a few who did poorly anyway).
d) Bimodal – writing ability in a university writing center (some would come because they need help, the rest would be skilled tutors, and there would be fewer in the middle) or income for a sample that includes people from the U.S. and people from a third world country (some would center on a very low income, and some around a much higher income).
e) Skewed to the right, with a mode and median of 0 but a positive mean – number of times students have eaten snake in their lives (most would never have eaten it, but a few would have tried it once, perhaps on Fear Factor, and an even smaller number would have had it several times) or number of times married for a sample of high school students (most would not have been married at all, but a few would have been married once, and an even smaller number would have been married more than once).

2.133 Political conservatism and liberalism

a) As seen in Example 12, one need not add up every separate number when calculating a mean. This would be unwieldy with the political conservatism and liberalism data. We would have to add up 69 ones, 240 twos, etc. (all the way up to 68 sevens), then divide by the 1933 people in the study. There's a far easier way. We can find the sum of all values in the study, $\sum x$, by multiplying each possible value (1–7 in this case) by its frequency.

$$\bar{x} = \frac{\sum x}{n} = \frac{69(1) + 240(2) + 221(3) + 740(4) + 268(5) + 327(6) + 68(7)}{1933} = \frac{7950}{1933} = 4.11$$

b) The mode, the most common score, is four.
c) The median would be the 967th score. In this case, that category is four.

2.134 Mode but not median and mean

We use the mode when we're interested in the category with the highest frequency, as opposed to merely finding the "center" of the data. To find a mean or median, we must have observations that measure a quantity. With unordered categories, observations do not do this. But, we can still find the most common outcome, so the mode is appropriate.

2.135 Multiple choice - GRE scores

The best answer is (a). 127 and 129 are close, but 649 is relatively larger than 529.

2.136 Multiple choice - Facts about s

The best answer is (b), *s* can be zero if all observations hold the same value.

2.137 Multiple choice - Relative GPA

The best answer is (a). The standard deviation would allow her to calculate her *z*-score.

2.138 True or false

a) False, consider the following data set: 3, 3, 3, 3, 3. Note that the mean = median = mode = 3.

b) False, consider the following data set: 1 2 3 4. The mean is 2.5 which is not one of the data points.

c) True, when *n* is odd, the median is the data point in the $\left(\frac{n+1}{2}\right)$th position of the sorted data.

d) True, by definition, the median is the second quartile which is also known as the 50th percentile.

2.139 Bad statistic

The standard deviation was incorrectly recorded. The standard deviation represents a typical scores distance from the mean. For grades ranging between 26 and 100, a standard deviation of 76 is way too large.

2.140 True or false: Soccer

False. The mean would be inflated by the wage bills of the few exorbitant rich clubs, but the magnitudes of these wage bills would not affect the median. Thus, the mean would be larger than the median.

♦♦2.141 Mean for grouped data

In Exercise 2.43 or 2.133, the mean could be expressed as a sum. Before, the mean was calculated by multiplying each score by its frequency, then summing these and dividing by the total number of subjects. Alternatively, we could first divide each frequency by the number of subjects, rather than dividing the sum by the number of subjects. Dividing the frequency for a given category by the total number of subjects would give us the proportion. We are just changing the order in which we perform the necessary operations to calculate the mean.

♦♦2.142 Worldwide airline fatalities

a) The median falls at the 50th percentile. In this case, the 50th percentile falls in the group that is 27 fatal accidents (61.5% observations are 27 or less but only 46.2% are 26 or less). Thus, the median is 27.

b) If the distribution is bell-shaped, the mean would fall in the middle, and be about 27. Further, the Empirical Rule would apply, and nearly all scores would fall within three standard deviations of the mean. If nearly all scores fall within 10 fatal accidents from the mean (17 is 10 fatal accidents below the mean of 27, and 37 is 10 fatal accidents above it), the standard deviation would be about 10 divided by 3, or about 3.3.

♦♦2.143 Range and standard deviation approximation

Based on the work of statisticians (the Empirical Rule), we know that most, if not all, data points fall within three standard deviations of the mean if we have a bell-shaped distribution. The formula for this is $\bar{x} \pm 3s$. If the region from three standard deviations below the mean to three standard deviations above the mean encompasses just about everyone in the data set, we could add the section below the mean (3*s*) to the section above the mean (3*s*) to get everyone in the data set. $3s + 3s = 6s$. Because the range is defined as everyone in the dataset, we can say that the range is equal, approximately, to 6*s*.

♦♦2.144 Range the least resistant

There are only two observations that are taken into account by the range, the minimum and maximum scores. The range is the difference between these two, so the range increases exactly the same amount as one of these scores increases (or in the case of the minimum, decreases). The mean and standard deviation, however, take the magnitude of all observations into account. Although an extreme score would pull the mean in its direction, and would increase the standard deviation, this "pull" would be offset, at least to some degree, by the values of the other observations.

♦♦2.145 Using MAD to measure variability

a) With greater variability, numbers tend to be further from the mean. Thus, the absolute values of their deviations from the mean would be larger. When we take the average of all these values, the overall MAD is larger than with distributions with less variability.

b) The MAD is more resistant than the standard deviation because by squaring the deviations using the standard deviation formula, a large deviation has greater effect.

♦♦2.146 Rescale the data

a) $c = 20$, new mean: $57 + 20 = 77$, standard deviation: 20 (unchanged)

b) $c = \frac{1}{2}$, new mean: \$39,000/2 = 19,500 pounds, standard deviation: \$15,000/2 = 7500 pounds

c) Linear transformations do not change the shape of the distribution.

Chapter Problems: Student Activities

💻2.147 The average student

Answers will vary.

💻2.148 Create own data.

Answers will vary.

💻2.149 GSS

Answers will vary.