

1.86 (a) 8.3% were elite soccer players; 5% developed arthritis of the hip or knee. **(b)** 14.1% developed arthritis; 23.3% were elite soccer players. **(c)** Of the elite soccer players, 14.1% developed arthritis later in life. Of the non-elite soccer players, 4.2% developed arthritis later in life. Of those that did not play, 4.2% developed arthritis later in life. The data do appear to confirm the suspicion that more serious soccer players were more likely to develop arthritis later in life.

PD Section 1.3 Overview

The video overview can be found by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

TRM Section 1.3 Alternate Examples

You can find the Alternate Examples for this section in Microsoft Word format by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

Making Connections

In earlier grades, most students learn mean, median, and mode together and assume that all of them are measures of center. The mode may not be a measure of center, as there is no guarantee that the mode is near the center of a distribution. The mode can be the smallest value in the data set.

Recycle and Review

86. Risks of playing soccer (1.1) A study in Sweden looked at former elite soccer players, people who had played soccer but not at the elite level, and people of the same age who did not play soccer. Here is a two-way table that classifies these individuals by whether or not they had arthritis of the hip or knee by their mid-fifties.³⁶

		Soccer level		
		Elite	Non-elite	Did not play
Whether person developed arthritis	Yes	10	9	24
	No	61	206	548

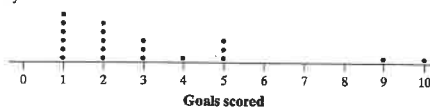
- (a) What percent of the people in this study were elite soccer players? What percent of the people in this study developed arthritis?
- (b) What percent of the elite soccer players developed arthritis? What percent of those who got arthritis were elite soccer players?
- (c) Researchers suspected that the more serious soccer players were more likely to develop arthritis later in life. Do the data confirm this suspicion? Calculate appropriate percentages to support your answer.

SECTION 1.3 Describing Quantitative Data with Numbers

LEARNING TARGETS By the end of the section, you should be able to:

- Calculate measures of center (mean, median) for a distribution of quantitative data.
- Calculate and interpret measures of variability (range, standard deviation, IQR) for a distribution of quantitative data.
- Explain how outliers and skewness affect measures of center and variability.
- Identify outliers using the $1.5 \times IQR$ rule.
- Make and interpret boxplots of quantitative data.
- Use boxplots and numerical summaries to compare distributions of quantitative data.

How much offense did the 2016 U.S. women's soccer team generate? The dot plot (reproduced from Section 1.2) shows the number of goals the team scored in 20 games played.



The *mode* of a data set is the most frequently occurring value. For the distribution of goals scored by the 2016 U.S. women's soccer team, the mode is 1. As you can see, the mode is rarely a good way to describe the center of the distribution.

The distribution is right-skewed and single-peaked (unimodal). The games in which the team scored 9 and 10 goals appear to be outliers. How can we describe the center and variability of this distribution?

Measuring Center: The Mean

The most common measure of center is the **mean**.

DEFINITION The mean

The **mean** of a distribution of quantitative data is the average of all the individual data values. To find the mean, add all the values and divide by the total number of data values.

If the n observations are $x_1, x_2, \dots, x_n$, the sample mean $\bar{x}$ (pronounced “x-bar”) is given by the formula

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x_i}{n}$$

The Σ (capital Greek letter sigma) in the formula is short for “add them all up.” The subscripts on the observations x_i are just a way of keeping the n data values distinct. They do not necessarily indicate order or any other special facts about the data.

EXAMPLE

How many goals? Calculating the mean



PROBLEM: Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women’s soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

- (a) Calculate the mean number of goals scored per game by the team. Show your work.
 (b) The earlier description of these data (page 35) suggests that the games in which the team scored 9 and 10 goals are possible outliers. Calculate the mean number of goals scored per game by the team in the other 18 games that season. What do you notice?

SOLUTION:

$$\begin{aligned} \text{(a)} \quad \bar{x} &= \frac{5 + 5 + 1 + 10 + 5 + 2 + 1 + 1 + 2 + 3 + 3 + 2 + 1 + 4 + 2 + 1 + 2 + 1 + 9 + 3}{20} \\ &= \frac{63}{20} = 3.15 \text{ goals} \end{aligned}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

(b) The mean for the other 18 games is

$$\begin{aligned} \bar{x} &= \frac{5 + 5 + 1 + 5 + 2 + 1 + 1 + 2 + 3 + 3 + 2 + 1 + 4 + 2 + 1 + 2 + 1 + 3}{18} \\ &= \frac{44}{18} = 2.44 \text{ goals} \end{aligned}$$

These two games increased the team’s mean number of goals scored per game by 0.71 goals.

FOR PRACTICE, TRY EXERCISE 87

The notation $\bar{x}$ refers to the mean of a *sample*. Most of the time, the data we encounter can be thought of as a sample from some larger population. When we need to refer to a *population mean*, we’ll use the symbol μ (Greek letter mu, pronounced “mew”). If you have the entire population of data available, then you calculate μ in just the way you’d expect: add the values of all the observations, and divide by the number of observations.

We call $\bar{x}$ a *statistic* and μ a *parameter*. Remember s and p: statistics come from samples and parameters come from populations.

AP® EXAM TIP

The formula for the sample mean $\bar{x}$ is included on the AP® Statistics formula sheet that can be used on both parts of the AP® Statistics exam. We recommend that you point out the Formula sheet in the appendix (page F-1) and give each student a copy right away so he or she can get accustomed to referring to it. Consider making laminated copies for students to use during tests and quizzes.

Preparing for Inference

In Chapter 4, we will reinforce that a value calculated from a sample (a statistic) is used to estimate a value for a population (a parameter). We use different notation for statistics than we do for parameters so readers will know if the reported value is an estimate or the true value.

ALTERNATE EXAMPLE

Skill 2.C

How many likes on Instagram for ASA? Calculating the mean

PROBLEM:

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews) to post updates on new statistical publications and adorable normal distribution plushies. Here are the number of Instagram likes for 10 posts selected at random:

16 4 8 7 8
6 15 2 9 5

SOLUTION:

$$\text{(a)} \quad \bar{x} = \frac{\sum x_i}{n} = \frac{16 + 4 + 8 + 7 + 8 + 6 + 15 + 2 + 9 + 5}{10} = \frac{80}{10} = 8 \text{ likes}$$

(b) The mean for the other 8 posts is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{4 + 8 + 7 + 8 + 6 + 2 + 9 + 5}{8} = \frac{49}{8} = 6.1 \text{ likes}$$

These two posts increased the mean number of Instagram likes per post by 1.9 likes.

- (a) Calculate the mean number of Instagram likes for these 10 posts. Show your work.
 (b) The posts with 15 and 16 likes are possible outliers. Calculate the mean number of Instagram likes in the other 8 posts. What do you notice?

Teaching Tip

Resistant = doesn't want to change



ACTIVITY OVERVIEW

To prepare for using this activity, watch the overview video by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

Time: 10 minutes

Materials: A foot-long ruler, pencil, and 5 pennies per student or group of students

Teaching Advice: Students can work in pairs or small groups if it is difficult to balance the pennies and keep the pencil from moving. This activity helps students recognize that the mean is the balancing point/center of mass/centroid of the distribution. Using pencils with hexagonal (not round) shafts will make this activity easier to complete

Answers:

1. The pencil is at the mean value, 6.
2. The other penny should balance the ruler when it is at 4. The mean is still 6.
3. The remaining two pennies should be at 8, for a total of three pennies. The mean is still 6.
4. Answers will vary. One possible answer is that the sum of the distances to the right of 6 will "balance out" the sum of the distances to the left of 6.

DEFINITION Statistic, Parameter

A **statistic** is a number that describes some characteristic of a sample.

A **parameter** is a number that describes some characteristic of a population.



The preceding example illustrates an important weakness of the mean as a measure of center: **the mean is sensitive to extreme values in a distribution.** These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. We say that the mean is not a **resistant** measure of center.

DEFINITION Resistant

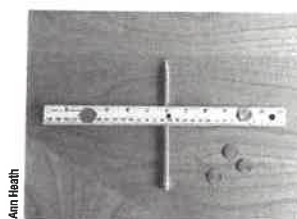
A statistical measure is **resistant** if it isn't sensitive to extreme values.

The mean of a distribution also has a physical interpretation, as the following activity shows.

ACTIVITY

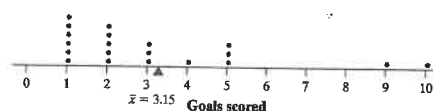
Mean as a "balance point"

In this activity, you'll investigate an important property of the mean.



1. Stack 5 pennies on top of the 6-inch mark on a 12-inch ruler. Place a pencil under the ruler to make a "seesaw" on a desk or table. Move the pencil until the ruler balances. What is the relationship between the location of the pencil and the mean of the five data values 6, 6, 6, 6, and 6?
2. Move one penny off the stack to the 8-inch mark on your ruler. Now move one other penny so that the ruler balances again without moving the pencil. Where did you put the other penny? What is the mean of the five data values represented by the pennies now?
3. Move one more penny off the stack to the 2-inch mark on your ruler. Now move both remaining pennies from the 6-inch mark so that the ruler still balances with the pencil in the same location. Is the mean of the data values still 6?
4. Discuss with your classmates: Why is the mean called the "balance point" of a distribution?

The activity gives a physical interpretation of the mean as the balance point of a distribution. For the data on goals scored in each of 20 games played by the 2016 U.S. women's soccer team, the dotplot balances at $\bar{x} = 3.15$ goals.



Teaching Tip

If you don't have the time to do the ruler activity, you can illustrate the basic concepts quickly on the board using dotplots. Instead of placing 5 pennies on the 6-inch mark, place a stack of 5 dots above 6 on a dotplot running from 0 to 12. Then "move" one of the dots from 6 to 8 and ask students what needs to happen for the mean to remain equal to 6. When they suggest "moving" a dot from 6 to 4, point out that the deviations from 6 (the mean) are in balance. Now "move" another dot from 6 to 2 and have students discuss how to keep the mean balanced at 6.

Measuring Center: The Median

We could also report the value in the “middle” of a distribution as its center. That’s the idea of the **median**.

DEFINITION Median

The **median** is the midpoint of a distribution, the number such that about half the observations are smaller and about half are larger.

To find the median, arrange the data values from smallest to largest.

- If the number n of data values is odd, the median is the middle value in the ordered list.
- If the number n of data values is even, use the average of the two middle values in the ordered list as the median.

The median is easy to find by hand for small sets of data. For instance, here are the data from Section 1.2 on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA:

22.4 22.4 22.3 23.3 22.3 22.3 22.5 22.4 22.1 21.5 22.0 22.2 22.7
22.8 22.4 22.6 22.9 22.5 22.1 22.4 22.2 22.9 22.6 21.9 22.4

Start by sorting the data values from smallest to largest:

21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

There are $n = 25$ data values (an odd number), so the median is the middle (13th) value in the ordered list, the bold 22.4.

EXAMPLE

How many goals? Finding the median

PROBLEM: Here are the data on the number of goals scored in 30 games played by the 2016 U.S. women’s soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Find the median.

SOLUTION:

1 1 1 1 1 2 2 2 **2** 3 3 3 4 5 5 5 9 10

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



To find the median, sort the data values from smallest to largest. Because there are $n = 20$ data values (an even number), the median is the average of the middle two values in the ordered list.

FOR PRACTICE, TRY EXERCISE 89

Teaching Tip

In *TPS 6e*, we do not use a symbol to represent the median the way we do for the mean ($\bar{x}$ or μ). Although some people use the symbol M (e.g., $M = 20$), we have found it clearer for students to simply write “median = 20.”

Making Connections

Many students have learned the “crossing out method” for finding a median, where they successively cross out values from the ends of the distribution, until they get to the center.

ALTERNATE EXAMPLE

Skill 2.C

How many likes on Instagram for ASA? Finding the median

PROBLEM:

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews). Here are the number of Instagram likes for 10 posts selected at random:

16 4 8 7 8
6 15 2 9 5

Find the median.

SOLUTION:

2 4 5 6 **7** **8** 8 9 15 16

The median is $\frac{7+8}{2} = 7.5$ likes.

Comparing the Mean and the Median

Which measure—the mean or the median—should we report for a distribution? That depends on both the shape of the distribution and whether there are any outliers.

- **Shape:** Figure 1.13 shows the mean and median for three different shapes. Notice how these two measures compare in each case. The mean is pulled in the direction of skewness in a skewed distribution.

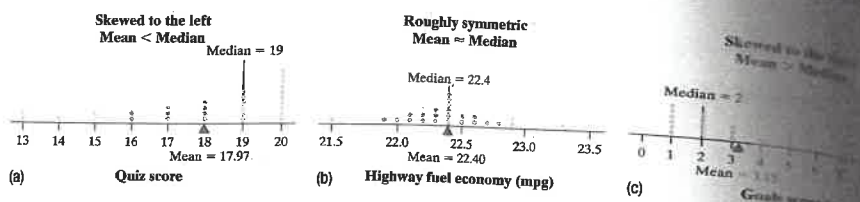


FIGURE 1.13 Dotplots that show the relationship between the mean and median in distributions with different shapes: (a) 30 statistics students on a 20-point quiz, (b) highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners, and (c) number of goals scored in 20 games played by the 2016 U.S. women's soccer team.

You can compare how the mean and median behave by using the *Mean and Median* applet at the book's website, highschool.bfwpub.com/updatedtps6e.

- **Outliers:** We noted earlier that the mean is sensitive to extreme values. If we remove the two possible outliers (9 and 10) in 1.13(c), the mean number of goals scored per game decreases from 3.15 to 2.44. The median number of goals scored is 2, which includes these two games or not. The median is a resistant measure of center, but the mean is not.

EFFECT OF SKEWNESS AND OUTLIERS ON MEASURES OF CENTER

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't. For a right-skewed distribution, we expect the mean to be greater than the median. For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't.

The mean and median measure center in different ways, and both are useful. In Major League Baseball (MLB), the distribution of player salaries is strongly skewed to the right. Most players earn close to the minimum salary (which was \$507,500 in 2016), while a few earn more than \$20 million. The median salary for MLB players in 2016 was about \$1.5 million—but the mean salary was about \$4.4 million. Clayton Kershaw, Miguel Cabrera, John Lester, and several other highly paid superstars pulled the mean up but that did not affect the median.

Teaching Tip

Some textbooks refer to values that are “sensitive” to outliers as being “nonresistant.”

Teaching Tip: Using Technology

Using the *Mean and Median* applet at the book's website, highschool.bfwpub.com/updatedtps6e, create a distribution that is fairly symmetric, where the mean and median are very close. Ask students what will happen if you add a few outliers on the far right. Then add the outliers and watch the mean move to the right (while the median remains relatively unchanged). Another option is to have students create three distributions (roughly symmetric, skewed right, and skewed left). Ask them to compare the mean and median for each distribution.

AP® EXAM TIP

If students are asked to choose between the mean and median as a measure of center, be sure they justify their choice based on the shape of the distribution and whether there are any possible outliers.

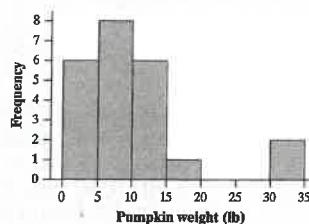
The median gives us a good idea of what a “typical” MLB salary is. If we want to know the total salary paid to MLB players in 2016, however, we would multiply the mean salary by the total number of players: (\$4.4 million)(862) $\approx$ \$3.8 billion!



CHECK YOUR UNDERSTANDING

Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here, along with a histogram of the data.

3.6 4.0 9.6 14.0 11.0 12.4 13.0 2.0 6.0 6.6 15.0 3.4
12.7 6.0 2.8 9.6 4.0 6.1 5.4 11.9 5.4 31.0 33.0



1. Calculate the mean weight of the pumpkins.
2. Find the median weight of the pumpkins.
3. Would you use the mean or the median to summarize the typical weight of a pumpkin in this contest? Explain.

Measuring Variability: The Range

Being able to describe the shape and center of a distribution is a great start. However, two distributions can have the same shape and center, but still look quite different.

Figure 1.14 shows comparative dotplots of the length (in millimeters) of separate random samples of PVC pipe from two suppliers, A and B.³⁷ Both distributions are roughly symmetric and single-peaked (unimodal), with centers at about 600 mm, but the variability of these two distributions is quite different. The sample of pipes from Supplier A has much more consistent lengths (less variability) than the sample from Supplier B.

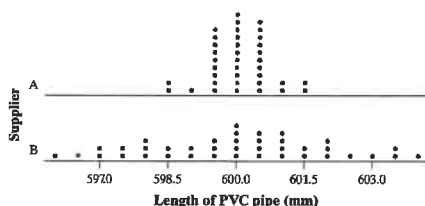


FIGURE 1.14 Comparative dotplots of the length of PVC pipes in separate random samples from Supplier A and Supplier B.

Teaching Tip

To introduce measures of variability, draw two dotplots with the same center and the same shape but very different amounts of variability. Hopefully, this will convince students that simply describing the shape and center does not give a complete picture of a distribution.

Answers to CYU

1. The mean weight of the pumpkins is $\bar{x} = \frac{3.6 + 4.0 + 9.6 + \dots + 5.4 + 31 + 33}{23} = 9.935$ pounds.
2. First we must put the weights in order: 2.0, 2.8, 3.4, 3.6, 4.0, 4.0, 5.4, 5.4, 6.0, 6.0, 6.1, **6.6**, 9.6, 9.6, 11.0, 11.9, 12.4, 12.7, 13.0, 14.0, 15.0, 31.0, and 33.0. Because there are 23 weights, the median is the 12th weight in this ordered list. The median weight of the pumpkins is 6.6 pounds.
3. I would use the median to summarize the typical weight of a pumpkin in this contest because the distribution of pumpkin weights is skewed to the right with two possible upper outliers: 31.0 and 33.0 pounds.

There are several ways to measure the variability of a distribution. The simplest is the **range**.

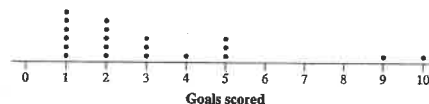
DEFINITION Range

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team, along with a dotplot:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3



Teaching Tip

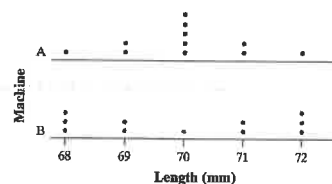
Insist that students provide a single number when discussing the range of a distribution. The range is the distance between the minimum and maximum value in the distribution.



The range of this distribution is $10 - 1 = 9$ goals. Note that the **range** of a data set is a **single number**. In everyday language, people sometimes say things like, "The data values range from 1 to 10." A correct statement is "The number of goals scored in 20 games played by the 2016 U.S. women's soccer team varies from 1 to 10, a range of 9 goals."

The range is *not* a resistant measure of variability. It depends on only the maximum and minimum values, which may be outliers. Look again at the data on goals scored by the 2016 U.S. women's soccer team. Without the possible outliers at 9 and 10 goals, the range of the distribution would decrease to $5 - 1 = 4$ goals.

The following graph illustrates another problem with the range as a measure of variability. The parallel dotplots show the lengths (in millimeters) of a sample of 11 nails produced by each of two machines.³⁸ Both distributions are centered at 70 mm and have a range of $72 - 68 = 4$ mm. But the lengths of the nails made by Machine B clearly vary more from the center of 70 mm than the nails made by Machine A.



Measuring Variability: The Standard Deviation

If we summarize the center of a distribution with the mean, then we should use the **standard deviation** to describe the variation of data values around the mean. To obtain the standard deviation, we start by calculating the **variance**.

Teaching Tip

It is not always possible to calculate the standard deviation if we do not have all the individual values in the data set (e.g., if we are given a histogram). In this case, the estimated range may be the best measure of variability.

Teaching Tip

The average of the absolute values of the deviations from the mean is called the **mean absolute deviation**, or **MAD**. This is not on the AP® Statistics exam, but you can tell students that it is a legitimate way to measure the variability of a distribution. However, it is not nearly as useful as the standard deviation for many reasons. For example, the Normal distributions are determined by their mean and standard deviation, not the mean absolute deviation.

AP® EXAM

The formula with the AP gives the standard deviation is

$$s_x = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}$$

The population is calculated as squared size N (population square root)

PRO friend

Calc

AP® EXAM TIP

The formula sheet provided with the AP® Statistics exam gives the sample standard deviation in the equivalent form

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

DEFINITION Standard deviation, Variance

The **standard deviation** measures the typical distance of the values in a distribution from the mean. It is calculated by finding an average of the squared deviations and then taking the square root. This average squared deviation is called the **variance**. If the values in a data set are $x_1, x_2, \dots, x_n$, the sample variance s_x^2 is given by the formula

$$s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

The sample standard deviation s_x is the square root of the variance:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

How do we calculate the standard deviation s_x of a quantitative data set with n values? Here are the steps.

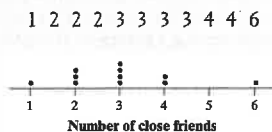
HOW TO CALCULATE THE SAMPLE STANDARD DEVIATION s_x

- Find the mean of the distribution.
- Calculate the *deviation* of each value from the mean:
deviation = value – mean
- Square each of the deviations.
- Add all the squared deviations, then divide by $n - 1$ to get the sample variance.
- Take the square root to return to the original units.

The notation s_x refers to the standard deviation of a *sample*. When we need to refer to the standard deviation of a *population*, we'll use the symbol σ (Greek lowercase sigma). We often use the sample statistic s_x to estimate the population parameter σ .

EXAMPLE**How many friends?
Calculating and interpreting standard deviation**

PROBLEM: Eleven high school students were asked how many “close” friends they have. Here are their responses, along with a dotplot:



Calculate the sample standard deviation. Interpret this value.



LanaBaker/Getty Images

Teaching Tip: AP® Connections

AP® Psychology students are taught to divide by n for the standard deviation calculation. This is not necessarily wrong, but it presupposes that the data given are the population data.

Teaching Tip

Why square each deviation? To make all values positive. *Why square root?* Because we squared each deviation in Step 3. We are undoing that process so we can return to the original unit of measurement. *Why divide by $n - 1$?* The standard deviation we are calculating is just an estimate of the true standard deviation—the “average” distance all the values in the population are from the mean of the population. Because we don’t know the mean of the population (μ), we can’t calculate the deviations from μ . So we use the mean of the sample ($\bar{x}$) instead. However, because $\bar{x}$ is the exact balancing point of the data, the data will almost always be closer to $\bar{x}$, on average, than they will be to μ . Thus, the sum of the squared deviations from $\bar{x}$ will underestimate the sum of the squared deviations from μ . To correct for this, we divide by a slightly smaller number when finding the average ($n - 1$ instead of n). We will demonstrate this with a simulation in Chapter 7.

Preparing for Inference

In Chapter 4, we will reinforce that a value calculated from a sample (a statistic) is used to estimate a value for a population (a parameter). We use different notation for statistics than we do for parameters, so readers will know if the reported value is an estimate or the true value.

AP® EXAM TIP

The equation for standard deviation is on the formula sheet provided on both parts of the AP® Statistics exam, but it is unlikely that students will be asked to perform this calculation. It is more important that they can interpret the standard deviation as the typical distance the values in a distribution are from the mean.

Teaching Tip

Going through a single example of calculating standard deviation by hand is worthwhile in helping students to understand the correct interpretation of standard deviation as the “typical distance from the mean” as well as to understand some of the properties of standard deviation.

SOLUTION:

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1-3=-2	$(-2)^2=4$
2	2-3=-1	$(-1)^2=1$
2	2-3=-1	$(-1)^2=1$
2	2-3=-1	$(-1)^2=1$
3	3-3=0	$0^2=0$
3	3-3=0	$0^2=0$
3	3-3=0	$0^2=0$
3	3-3=0	$0^2=0$
4	4-3=1	$1^2=1$
4	4-3=1	$1^2=1$
6	6-3=3	$3^2=9$
		Sum=18

$$s_x^2 = \frac{18}{11-1} = 1.80$$

$$s_x = \sqrt{1.80} = 1.34 \text{ close friends}$$

Interpretation: The number of close friends these students have typically varies by about 1.34 close friends from the mean of 3 close friends.

To calculate the sample standard deviation:

- Find the mean of the distribution.
- Calculate the deviation of each value from the mean:
deviation = value - mean
- Square each of the deviations.
- Add all the squared deviations, then divide by $n - 1$ to get the sample variance.
- Take the square root to return to the original units.

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

FOR PRACTICE, TRY EXERCISE 29

ALTERNATE EXAMPLE

Skill 2.C

How many likes on Instagram for ASA?
Calculating and interpreting standard deviation

PROBLEM:

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews). Here are the number of Instagram likes for 10 posts selected at random:

2 4 5 6 7
8 8 9 15 16

Calculate the sample standard deviation. Interpret this value in context.

SOLUTION:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+4+5+6+7+8+8+9+15+16}{10} = \frac{80}{10} = 8 \text{ likes}$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
2	2-8=-6	$(-6)^2=36$
4	4-8=-4	$(-4)^2=16$
5	5-8=-3	$(-3)^2=9$
6	6-8=-2	$(-2)^2=4$
7	7-8=-1	$(-1)^2=1$
8	8-8=0	$0^2=0$
8	8-8=0	$0^2=0$
9	9-8=1	$1^2=1$
15	15-8=7	$7^2=49$
16	16-8=8	$8^2=64$

Sum = 180

In the preceding example, the sample variance is $s_x^2 = 1.80$. Unfortunately, the units are “squared close friends.” Because variance is measured in squared units, it is not a very helpful way to describe the variability of a distribution.

Think About It

WHY IS THE STANDARD DEVIATION CALCULATED IN SUCH A COMPLEX WAY? Add the deviations from the mean in the preceding example. You should get a sum of 0. Why? Because the mean is the balance point of the distribution. We square the deviations to avoid the positive and negative deviations balancing each other out and adding to 0. It might seem strange to “average” the squared deviations by dividing by $n - 1$. We’ll explain the reason for doing this in Chapter 7. It’s easier to understand why we take the square root to return to the original units (close friends).

$$s_x = \sqrt{\frac{180}{10-1}} = 4.47 \text{ likes}$$

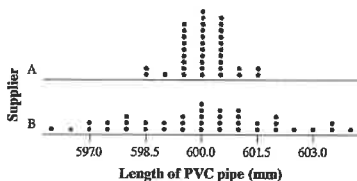
Interpretation: The number of Instagram likes for each ASA post typically varies by about 4.47 likes from the mean of 8 likes.

Making Connections

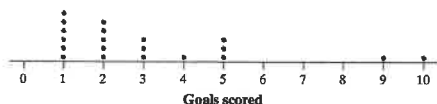
Why do we bother with variance? You will find out in Chapter 6 when we combine random variables and the variances will add up (not the standard deviations!).

More important than the details of calculating s_x are the properties of the standard deviation as a measure of variability:

- s_x is always greater than or equal to 0. $s_x = 0$ only when there is no variability, that is, when all values in a distribution are the same.
- Larger values of s_x indicate greater variation from the mean of a distribution. The comparative dotplot shows the lengths of PVC pipe in random samples from two different suppliers. Supplier A's pipe lengths have a standard deviation of 0.681 mm, while Supplier B's pipe lengths have a standard deviation of 2.02 mm. The lengths of pipes from Supplier B are typically farther from the mean than the lengths of pipes from Supplier A.



- s_x is not a resistant measure of variability. The use of squared deviations makes s_x even more sensitive than $\bar{x}$ to extreme values in a distribution. For example, the standard deviation of the number of goals scored in 20 games played by the 2016 U.S. women's soccer team is 2.58 goals. If we omit the possible outliers of 9 and 10 goals, the standard deviation drops to 1.46 goals.



- s_x measures variation about the mean. It should be used only when the mean is chosen as the measure of center.

In the close friends example, 11 high school students had an average of $\bar{x} = 3$ close friends with a standard deviation of $s_x = 1.34$. What if a 12th high school student was added to the sample who had 3 close friends? The mean number of close friends in the sample would still be $\bar{x} = 3$. How would s_x be affected? Because the standard deviation measures the typical distance of the values in a distribution from the mean, s_x would *decrease* because this 12th value is at a distance of 0 from the mean. In fact, the new standard deviation would be

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{18}{12-1}} = 1.28 \text{ close friends}$$

Measuring Variability: The Interquartile Range (IQR)

We can avoid the impact of extreme values on our measure of variability by focusing on the middle of the distribution. Start by ordering the data values from smallest to largest. Then find the **quartiles**, the values that divide the distribution into four

Teaching Tip

Use the formula for standard deviation

$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$ to help explain each of these properties.

Teaching Tip

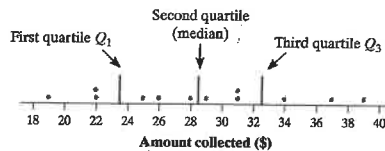
Point out that the definition for quartiles in this book implies that the median is never included when computing quartiles.

Teaching Tip: Using Technology

Statistical software packages may differ in how they calculate quartiles. However, the differences are usually small, especially when there is a large number of individuals in the data set.

Teaching Tip

Just as with range, be careful to identify *IQR* as a single value and not an interval of values. For example, the "*IQR* is 5 pairs of shoes" is correct, while the "*IQR* is from 6.5 to 11.5 shoes" is not correct.



groups of roughly equal size. The first quartile Q_1 lies one-quarter of the way up the list. The second quartile is the median, which is halfway up the list. The third quartile Q_3 lies three-quarters of the way up the list. The first and third quartiles mark out the middle half of the distribution.

For example, here are the amounts collected each hour by a charity at a local store: \$19, \$26, \$25, \$37, \$31, \$28, \$22, \$22, \$29, \$34, \$39, and \$31. The dotplot displays the data. Because there are 12 data values, the quartiles divide the distribution into 4 groups of 3 values.

DEFINITION Quartiles, First quartile Q_1 , Third quartile Q_3

The quartiles of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The first quartile Q_1 is the median of the data values that are to the left of the median in the ordered list.

The third quartile Q_3 is the median of the data values that are to the right of the median in the ordered list.

The interquartile range (*IQR*) measures the variability in the middle half of the distribution.

DEFINITION Interquartile range (*IQR*)

The interquartile range (*IQR*) is the distance between the first and third quartiles of a distribution. In symbols:

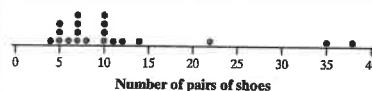
$$IQR = Q_3 - Q_1$$

Notice that the *IQR* is simply the range of the "middle half" of the distribution.

EXAMPLE**Boys and their shoes?**
Finding the *IQR*

PROBLEM: How many pairs of shoes does a typical teenage boy own? To find out, two AP® Statistics students surveyed a random sample of 20 male students from their large high school and recorded the number of pairs of shoes that each boy owned. Here are the data, along with a dotplot:

14 7 6 5 12 38 8 7 10 10 10 11 4 5 22 7 5 10 35 7



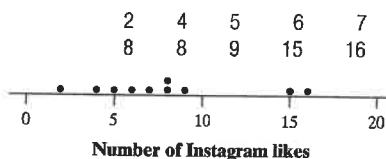
Peter Dinkley/Photodisc

ALTERNATE EXAMPLE**How many Instagram likes for ASA? Finding the *IQR***

Skill 2.C

PROBLEM:

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews). Here are the number of Instagram likes for 10 of their posts selected at random, along with a dotplot:

**SOLUTION:**

2 4 5 6 7 | 8 8 9 15 16
Median = 7.5

2 4 5 6 7 | 8 8 9 15 16

$Q_1 = 5$ $Q_3 = 9$
 $IQR = 9 - 5 = 4$ likes

Find the interquartile range.

Find the interquartile range.

SOLUTION:

4 5 5 5 6 7 7 7 7 8 10 10 10 11 12 14 22 35 38
Median = 9

Sort the data values from smallest to largest and find the median.

4 5 5 6 7 7 7 7 8 10 10 10 10 11 12 14 22 35 38
 $Q_1 = 6.5$ Median $Q_3 = 11.5$

Find the first quartile Q_1 and the third quartile Q_3 .

$IQR = 11.5 - 6.5 = 5$ pairs of shoes

$IQR = Q_3 - Q_1$

FOR PRACTICE, TRY EXERCISE 105

The quartiles and the interquartile range are *resistant* because they are not affected by a few extreme values. For the shoe data, Q_3 would still be 11.5 and the IQR would still be 5 if the maximum were 58 rather than 38.

Be sure to leave out the median when you locate the quartiles. In the preceding example, the median was not one of the data values. For the earlier close friends data set, we ignore the circled median of 3 when finding Q_1 and Q_3 .

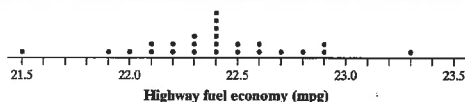
1 2 ② 2 3 ③ 3 3 ④ 4 6
 Q_1 Median Q_3



CHECK YOUR UNDERSTANDING

Here are data on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA, along with a dotplot:

22.4 22.4 22.3 23.3 22.3 22.3 22.5 22.4 22.1 21.5 22.0 22.2 22.7
22.8 22.4 22.6 22.9 22.5 22.1 22.4 22.2 22.9 22.6 21.9 22.4



- Find the range of the distribution.
- The mean and standard deviation of the distribution are 22.404 mpg and 0.363 mpg, respectively. Interpret the standard deviation.
- Find the interquartile range of the distribution.
- Which measure of variability would you choose to describe the distribution? Explain.

Numerical Summaries with Technology

Graphing calculators and computer software will calculate numerical summaries for you. Using technology to perform calculations will allow you to focus on choosing the right methods and interpreting your results.

Teaching Tip

Ask students to rank the standard deviation, IQR , and range from the most resistant (IQR) to the least resistant (range). Make sure they explain their reasoning!

Teaching Tip

Point out that the range is not a resistant measure of variability because the minimum and maximum values could be outliers. The IQR is a resistant measure of variability because it ignores the smallest 25% and the largest 25% of the values in a distribution.

✓ Answers to CYU

1. Range = Max - Min = 23.3 - 21.5 = 1.8 mpg

2. The standard deviation of 0.363 mpg tells us that the highway fuel economy of these 25 model year 2018 Toyota 4Runners typically varies by about 0.363 mpg from the mean of 22.404 mpg.

3. $Q_1 = \frac{22.2 + 22.2}{2} = 22.2$;

$Q_3 = \frac{22.6 + 22.6}{2} = 22.6$;

$IQR = Q_3 - Q_1 = 22.6 - 22.2 = 0.4$ mpg

4. I would use the IQR to describe variability because there appear to be one upper outlier and one lower outlier. The interquartile range is resistant to outliers, but the range and standard deviation are not.

Teaching Tip: Using Technology

We recommend that you start allowing students to calculate numerical summaries with technology at this point. Students will be expected to use their calculators on the AP® Statistics exam, which allows the writers of the AP® exam to ask more conceptual questions about the application and interpretation of the numerical summaries. Make sure your students get plenty of practice with these types of questions.

Making Connections

There is no frequency list because we want each entry in L1 to be used once. In Chapter 6, students will encounter distributions where each value is weighted according to a probability. In this scenario, we will use another list in the calculator to show the probabilities (or relative frequencies).

Teaching Tip

“Ruler” as in a measuring tool. Not a king.

3. Technology Corner COMPUTING NUMERICAL SUMMARIES

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

Let's find numerical summaries for the boys' shoes data from the example on page 64. We'll start by showing you how to compute summary statistics on the TI-83/84 and then look at output from computer software.

I. One-variable statistics on the TI-83/84

1. Enter the data in list L1.
2. Find the summary statistics for the shoe data.
 - Press **[STAT]** (CALC); choose 1-Var Stats. OS 2.55 or later: In the dialog box, press **[2nd]** **[1]** (L1) and **[ENTER]** to specify L1 as the List. Leave FreqList blank. Arrow down to Calculate and press **[ENTER]**. Older OS: Press **[2nd]** **[1]** (L1) and **[ENTER]**.
 - Press **[↓]** to see the rest of the one-variable statistics.

II. Output from statistical software We used Minitab statistical software to calculate descriptive statistics for the boys' shoes data. Minitab allows you to choose which numerical summaries are included in the output.

Descriptive Statistics: Shoes

Variable	N	Mean	StDev	Minimum	Q_1	Median	Q_3	Maximum
Shoes	20	11.65	9.42	4.00	6.25	9.00	11.75	38.00

Note: The TI-83/84 gives the first and third quartiles of the boys' shoes distribution as $Q_1 = 6.5$ and $Q_3 = 11.5$. Minitab reports that $Q_1 = 6.25$ and $Q_3 = 11.75$. What happened? Minitab and some other software use slightly different rules for locating quartiles. Results from the various rules are usually close to each other. Be aware of possible differences when calculating quartiles as they may affect more than just the IQR.

NORMAL FLOAT AUTO REAL RADIAN MP	
1-Var Stats	
$\bar{x}$	=11.65
Σx	=233
Σx^2	=4401
Sx	=9.421559822
σx	=9.183000599
n	=20
minX	=4
Q_1	=6.5

NORMAL FLOAT AUTO REAL RADIAN MP	
1-Var Stats	
$\bar{x}$	=11.65
Σx	=233
Σx^2	=4401
Sx	=9.421559822
σx	=9.183000599
n	=20
minX	=4
Q_1	=6.5
Med	=9
Q_3	=11.5
maxX	=38

Identifying Outliers

Besides serving as a measure of variability, the interquartile range (IQR) is used as a “ruler” for identifying outliers.

HOW TO IDENTIFY OUTLIERS: THE $1.5 \times IQR$ RULE

Call an observation an outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile. That is,

$$\text{Low outliers} < Q_1 - 1.5 \times IQR \quad \text{High outliers} > Q_3 + 1.5 \times IQR$$

Here are sorted data on the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA, along with a dotplot:

21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

There are other rules for determining outliers, such as “any value that is more than 2 (or 3) standard deviations from the mean.” We always use the $1.5 \times IQR$ rule in this book because it is based on statistics that are resistant to outliers, unlike the mean and standard deviation.

AP® EXAM

You may be asked whether a question has any outliers to state and to identify outliers.

EX

PROBLEM
in 20 years with a dotplot.

Identify an

SOLUTION

1 1 1 1

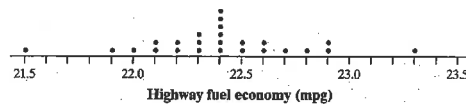
 $IQR = Q_3 - Q_1$

Low outlier

High outlier

There are

which the



Does the $1.5 \times IQR$ rule identify any outliers in this distribution? If you did the preceding Check Your Understanding, you should have found that $Q_1 = 22.2$ mpg, $Q_3 = 22.6$ mpg, and $IQR = 0.4$ mpg. For these data,

$$\text{High outliers} > Q_3 + 1.5 \times IQR = 22.6 + 1.5 \times 0.4 = 23.2$$

and

$$\text{Low outliers} < Q_1 - 1.5 \times IQR = 22.2 - 1.5 \times 0.4 = 21.6$$

The cars with estimated highway fuel economy ratings of 21.5 and 23.3 are identified as outliers.

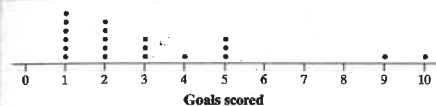
AP® EXAM TIP

You may be asked to determine whether a quantitative data set has any outliers. Be prepared to state and use the rule for identifying outliers.

EXAMPLE**How many goals?
Identifying outliers**

PROBLEM: Here are sorted data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team, along with a dotplot:

1 1 1 1 1 2 2 2 2 2 3 3 3 4 5 5 5 9 10



Identify any outliers in the distribution. Show your work.

SOLUTION:

1 1 1 1 1 2 2 2 2 2 3 3 3 4 5 5 5 9 10
 $Q_1 = 1$ Median = 2 $Q_3 = 4.5$

$$IQR = Q_3 - Q_1 = 4.5 - 1 = 3.5$$

$$\text{Low outliers} < 1 - 1.5 \times 3.5 = -4.25$$

$$\text{High outliers} > 4.5 + 1.5 \times 3.5 = 9.75$$

There are no data values less than -4.25 , but the game in which the team scored 10 goals is an outlier.

Low outliers $< Q_1 - 1.5 \times IQR$
 High outliers $> Q_3 + 1.5 \times IQR$

The game in which the team scored 9 goals is not identified as an outlier by the $1.5 \times IQR$ rule.

FOR PRACTICE, TRY EXERCISE 107



Icon Sports/WireImage/Getty Images

ALTERNATE EXAMPLE

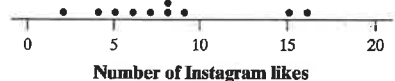
Skill 4.B

**How many likes on Instagram for ASA?
Identifying outliers**

PROBLEM:

The American Statistical Association (www.amstat.org) has an Instagram account (@amstatnews). Here are the number of Instagram likes for 10 of their posts selected at random, along with a dotplot:

2 4 5 6 7
 8 8 9 15 16



Identify any outliers in the distribution. Show your work.

SOLUTION:

2 4 5 6 7 8 8 9 15 16

$$Q_1 = 5 \quad \text{Median} = 7.5 \quad Q_3 = 9$$

$$IQR = 9 - 5 = 4$$

$$\text{Low outliers} < Q_1 - 1.5 \times IQR$$

$$= 5 - 1.5 \times 4 = -1$$

$$\text{High outliers} > Q_3 + 1.5 \times IQR$$

$$= 9 + 1.5 \times 4 = 15$$

There are no data values less than -1 , but the post that has 16 likes is an outlier ($16 > 15$). Note that 15 is *not* an outlier because $15 = Q_3 + 1.5 \times IQR$, and a high outlier must be *greater than* $Q_3 + 1.5 \times IQR$.

Teaching Tip

Reinforce the fact that the five-number summary divides the data roughly into quarters by asking students to approximate what percent of the data fall above or below certain values. For example, pick a value above the median but below Q_3 and ask students what percent of values in the distribution fall above the chosen value. They should say, "Somewhere between 25% and 50% of the values," unless they have the raw data.

Teaching Tip

The rules for determining outliers can quickly become a memorized formula (which often leads to student errors and weak long-term retention). Use the picture on this page to help students visualize going $1.5 \times IQR$ to the left of Q_1 and $1.5 \times IQR$ to the right of Q_3 . Even better, search for "IQR dance" on YouTube to find author Josh Tabor get his groove on.

A boxplot is sometimes called a *box-and-whisker plot*.

It is important to identify outliers in a distribution for several reasons:

1. **They might be inaccurate data values.** Maybe someone recorded a value as 10.1 instead of 101. Perhaps a measuring device broke down. Or maybe someone gave a silly response, like the student in a class survey who claimed to study 30,000 minutes per night! Try to correct errors like these if possible. If you can't, give summary statistics with and without the outlier.
2. **They can indicate a remarkable occurrence.** For example, in a graph of net worth, Bill Gates is likely to be an outlier.
3. **They can heavily influence the values of some summary statistics,** like the mean, range, and standard deviation.

Making and Interpreting Boxplots

You can use a dotplot, stemplot, or histogram to display the distribution of a quantitative variable. Another graphical option for quantitative data is a boxplot. A boxplot summarizes a distribution by displaying the location of 5 important values within the distribution, known as its **five-number summary**.

DEFINITION Five-number summary, Boxplot

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1 , the median, the third quartile Q_3 , and the maximum. A **boxplot** is a visual representation of the five-number summary.

Figure 1.15 illustrates the process of making a boxplot. The dotplot in Figure 1.15(a) shows the data on EPA estimated highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners. We have marked the first quartile, the median, and the third quartile with vertical green lines. The process of testing for outliers with the $1.5 \times IQR$ rule is shown in red. Because the values of 21.5 mpg and 23.3 mpg are outliers, we mark these separately. To get the finished boxplot in Figure 1.15(b), we make a box spanning from Q_1 to Q_3 and then draw "whiskers" to the smallest and largest data values that are not outliers.

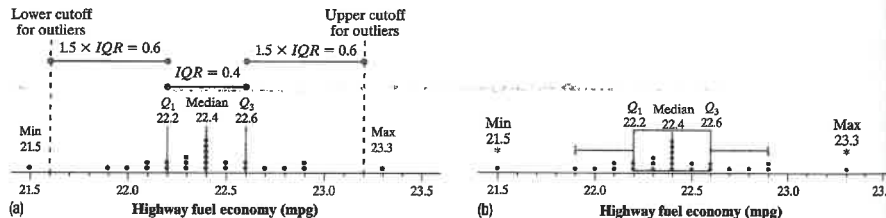


FIGURE 1.15 A visual illustration of how to make a boxplot for the Toyota 4Runner highway gas mileage data. (a) Dotplot of the data with the five-number summary and $1.5 \times IQR$ marked. (b) Boxplot of the data with outliers identified (*).

As you can see, it is fairly easy to make a boxplot by hand for small sets of data. Here's a summary of the steps.



COMMON STUDENT ERROR

Many students refer to the box in a boxplot as the *IQR*. For example, they will say "the median is in the middle of the *IQR*." This is incorrect because the *IQR* is a number (the length of the box), not the box itself.



COMMON STUDENT ERROR

When describing the shape of a boxplot, many students will describe a symmetric boxplot as "approximately Normal." This is incorrect because a boxplot does not reveal the location of the modes or any gaps for a distribution. It is sometimes difficult to assess the shape of a distribution using only a boxplot, because boxplots don't show peaks, gaps, or clusters. See 2019 free response #1 for a good example of this idea.

PROB
the co
shown

(a) M
(b) Ex
the

SOLU

(a)

Min
2.0 2
9.6 9

$IQR = c$
Low ou
High ou
The pu

ALTER
How m

PROBLE
Bedford,
.com/cal
an Insta
are the n
selected

HOW TO MAKE A BOXPLOT

- Find the five-number summary for the distribution.
- Identify outliers using the $1.5 \times IQR$ rule.
- Draw and label the axis. Draw a horizontal axis and put the name of the quantitative variable underneath, including units if applicable.
- Scale the axis. Look at the minimum and maximum values in the data set. Start the horizontal axis at a convenient number equal to or below the minimum and place tick marks at equal intervals until you equal or exceed the maximum.
- Draw a box that spans from the first quartile (Q_1) to the third quartile (Q_3).
- Mark the median with a vertical line segment that's the same height as the box.
- Draw whiskers—lines that extend from the ends of the box to the smallest and largest data values that are *not* outliers. Mark any outliers with a special symbol such as an asterisk (*).

We see from the boxplot in Figure 1.15 that the distribution of highway gas mileage ratings for this sample of model year 2018 Toyota 4Runners is roughly symmetric with one high outlier and one low outlier.

EXAMPLE

Picking pumpkins Making and interpreting boxplots

PROBLEM: Some students purchased pumpkins for a carving contest. Before the contest began, they weighed the pumpkins. The weights in pounds are shown here.

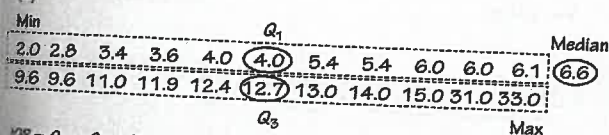
3.6 4.0 9.6 14.0 11.0 12.4 13.0 2.0 6.0 6.6 15.0 3.4
12.7 6.0 2.8 9.6 4.0 6.1 5.4 11.9 5.4 31.0 33.0

(a) Make a boxplot of the data.

(b) Explain why the median and IQR would be a better choice for summarizing the center and variability of the distribution of pumpkin weights than the mean and standard deviation.

SOLUTION:

(a)



$$IQR = Q_3 - Q_1 = 12.7 - 4.0 = 8.7$$

$$\text{Low outliers} < Q_1 - 1.5 \times IQR = 4.0 - 1.5 \times 8.7 = -9.05$$

$$\text{High outliers} > Q_3 + 1.5 \times IQR = 12.7 + 1.5 \times 8.7 = 25.75$$

The pumpkins that weighed 31.0 and 33.0 pounds are outliers.

To make the boxplot:

- Find the five-number summary.
- Identify outliers.
- Draw and label the axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers to the smallest and largest data values that are *not* outliers. Mark outliers with an asterisk.



Teaching Tip: Using Technology

Go to the Student Site and choose "One Quantitative Variable" from the Extra Applets. Type in the variable name and the data values. Click "Begin Analysis." Choose the graph option for boxplot.

AP® EXAM TIP

Some textbooks and the TI graphing calculators make the distinction between a boxplot that doesn't show outliers and a boxplot that shows outliers with a special symbol, such as an asterisk. Students can assume that any boxplot they encounter on the AP® Statistics exam displays outliers (if there are outliers).

SOLUTION:

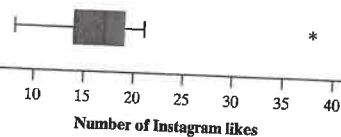
8 11 14 15 16 18 18 19 21 38

$$IQR = Q_3 - Q_1 = 19 - 14 = 5$$

$$\text{Outliers} < Q_1 - 1.5 \times IQR = 14 - 1.5 \times 5 = 6.5$$

$$\text{Outliers} > Q_3 + 1.5 \times IQR = 19 + 1.5 \times 5 = 26.5$$

The post that has 38 likes is an outlier ($38 > 26.5$).



(b) The distribution of number of likes on Instagram has one high outlier. Because the mean and standard deviation are sensitive to outliers, it would be better to use the resistant median and IQR .

ALTERNATE EXAMPLE

How many likes on Instagram for BFW? Making and interpreting boxplots

Skill 2.B

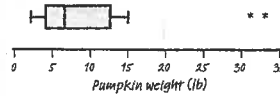
PROBLEM:

Bedford, Freeman & Worth (highschool.bfwpub.com/catalog), the publisher of this textbook, has an Instagram account (@bfbhighschool). Here are the number of Instagram likes for 10 posts selected at random:

19 21 18 14 15
18 8 16 11 38

(a) Make a boxplot of the data.

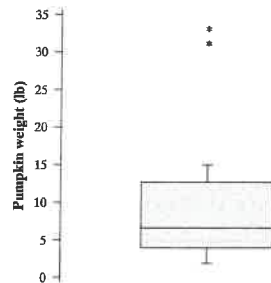
(b) Explain why the median and IQR would be a better choice for summarizing the center and variability of the distribution of number of Instagram likes than the mean and standard deviation.



(b) The distribution of pumpkin weights is skewed to the right with two high outliers. Because the mean and standard deviation are sensitive to outliers, it would be better to use the median and *IQR*, which are resistant.

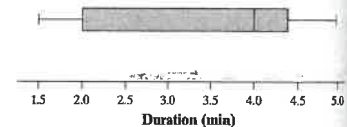
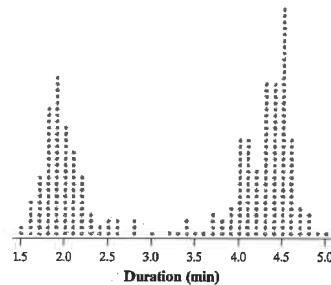
We know the distribution is skewed to the right because the left half of the distribution varies from 2.0 to 6.6 pounds, while the right half of the distribution (excluding outliers) varies from 6.6 to 15.0 pounds.

FOR PRACTICE, TRY EXERCISE 111



Boxplots provide a quick summary of the center and variability of a distribution. The median is displayed as a line in the central box, the interquartile range is the length of the box, and the range is the length of the entire plot, including outliers. Note that some statistical software orients boxplots vertically. At left is a vertical boxplot of the pumpkin weight data from the preceding example. You can see that the graph is skewed toward the larger values.

Boxplots do not display each individual value in a distribution. And boxplots don't show gaps, clusters, or peaks. For instance, the dotplot below left displays the duration, in minutes, of 220 eruptions of the Old Faithful geyser. The distribution of eruption durations is clearly double-peaked (*bimodal*). But a boxplot of the data hides this important information about the shape of the distribution.



Teaching Tip

In all other graphs, bigger areas correspond to higher frequencies. This isn't true with boxplots. For the duration data, the left side of the box is much longer than the right side of the box, but it contains an equal number of data values.

Teaching Tip

At this point, we have four displays for quantitative data: dotplots, stemplots, histograms, and boxplots. Ask students to identify the advantages and disadvantages of each.



CHECK YOUR UNDERSTANDING

Ryan and Brent were curious about the amount of french fries they would get in a large order from their favorite fast-food restaurant, Burger King. They went to several different Burger King locations over a series of days and ordered a total of 14 large fries. The weight of each order (in grams) is as follows:

PROB
ers for
gave e
The a
here, a
Apple
Samst

Apple
Samsu
Comp

ALTE
Who h

PROBLE
The Ame
(@amst
this text
are the r
from eac
summar

ASA
BFW

Group nam
1: ASA
2: BFW
Compare

165 163 160 159 166 152 166 168 173 171 168 167 170 170

1. Make a boxplot to display the data.
2. According to a nutrition website, Burger King's large fries weigh 160 grams, on average. Ryan and Brent suspect that their local Burger King restaurants may be skimping on fries. Does the boxplot in Question 1 support their suspicion? Explain why or why not.

Comparing Distributions with Boxplots

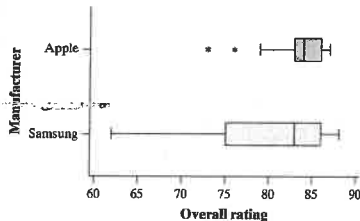
Boxplots are especially effective for comparing the distribution of a quantitative variable in two or more groups, as seen in the following example.

EXAMPLE

Which company makes better tablets? Comparing distributions with boxplots

PROBLEM: In a recent year, *Consumer Reports* rated many tablet computers for performance and quality. Based on several variables, the magazine gave each tablet an overall rating, where higher scores indicate better ratings. The overall ratings of the tablets produced by Apple and Samsung are given here, along with parallel boxplots and numerical summaries of the data.³⁹

Apple	87	87	87	87	86	86	86	86	84	84
	84	84	83	83	83	83	81	79	76	73
Samsung	88	87	87	86	86	86	86	84	84	83
	77	76	76	75	75	75	75	74	71	62



	$\bar{x}$	s_x	Min	Q_1	Median	Q_3	Max	IQR
Apple	83.45	3.762	73	83	84	86	87	3
Samsung	79.87	6.74	62	75	83	86	88	11

Compare the distributions of overall rating for Apple and Samsung.

Answers to CYU

$$1. M = \frac{166 + 167}{2} = 166.5 \text{ grams;}$$

$$Q_1 = 163 \text{ grams; } Q_3 = 170 \text{ grams.}$$

$$\text{The } IQR = 170 - 163 = 7 \text{ grams.}$$

An outlier is any value below

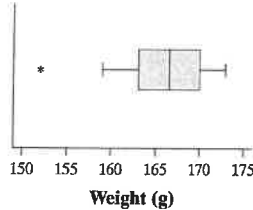
$$Q_1 - 1.5 \times IQR = 163 - 1.5 \times 7 = 152.5$$

grams or above

$$Q_3 + 1.5 \times IQR = 170 + 1.5 \times 7 = 180.5$$

grams. This means that the value

152 grams is an outlier. The boxplot is displayed below.



2. No, the graph does not support their suspicion. The first quartile of the distribution of weight is 163 grams, which means that at least 75% of the large fries that they purchased weighed more than the advertised weight of 160 grams.

ALTERNATE EXAMPLE

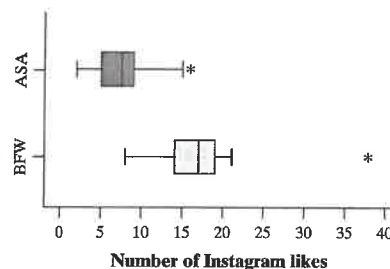
Skill 2.D

Who has more likes on Instagram? Comparing distributions with boxplots

PROBLEM:

The American Statistical Association (ASA) has an Instagram account (@amstatnews). Bedford, Freeman & Worth (BFW), the publisher of this textbook, also has an Instagram account (@bfwhighschool). Here are the number of Instagram likes for 10 randomly selected posts from each account, along with parallel boxplots and numerical summaries of the data:

ASA	16	4	8	7	8	6	15	2	9	5
BFW	19	21	18	14	15	18	8	16	11	38
Group name	n	mean	SD	min	Q_1	med	Q_3	max		
1: ASA	10	8	4.472	2	5	7.5	9	16		
2: BFW	10	17.8	8.08	8	14	17	19	38		



Compare the distributions of number of Instagram likes for ASA and BFW.

(continues)

ALTERNATE EXAMPLE (continued)

SOLUTION:

Shape: Both distributions of number of Instagram likes are slightly skewed to the right.

Outliers: There is one high outlier in the ASA distribution: the post with 16 likes. There is also one high outlier in the BFW distribution: the post with 38 likes.

Center: The BFW Instagram account had a higher median number of likes (17) than the ASA Instagram account (7.5). More importantly, 100% of the BFW posts had a number of likes greater than the median for the ASA account.

Variability: There is more variation in number of likes among the BFW posts than the ASA posts. The *IQR* for the BFW posts (5) is larger than the *IQR* for the ASA posts (4).

TRM Do You Know Your Geography?

Consider doing the activity, “Do you know your geography?,” found in the Teacher’s Resource Materials. The activity has each student answer a couple of questions about geography. However, there are two versions of the “geography quiz.” Comparing the results with parallel boxplots demonstrates how the results of a survey can be manipulated by changing the way the questions are asked. Thanks to Sanderson Smith for sharing this great activity.

TRM The Memory Game

Consider doing “The memory game” activity found in the Teacher’s Resource Materials. The activity has each student listen to some statements and then take a quiz about how much he or she remembers. However, one version of the quiz provides additional instructions. Comparing the results with parallel boxplots demonstrates how the results of a study can differ when given additional directions. Thanks to Adam Shrager for sharing this great activity.

SOLUTION:

Shape: Both distributions of overall ratings are skewed to the left.

Outliers: There are two low outliers in the Apple tablet distribution: overall ratings of 73 and 76. The Samsung tablet distribution has no outliers.

Center: The Apple tablets had a slightly higher median overall rating (84) than the Samsung tablets (83). More importantly, about 75% of the Apple tablets had overall ratings that were greater than or equal to the median for the Samsung tablets.

Variability: There is much more variation in overall rating among the Samsung tablets than the Apple tablets. The *IQR* for Samsung tablets (11) is almost four times larger than the *IQR* for Apple tablets (3).

Remember to compare shape, outliers, center, and variability!

Because of the strong skewness and outliers, use the median and *IQR* instead of the mean and standard deviation when comparing center and variability.

FOR PRACTICE, TRY EXERCISE 115

AP® EXAM TIP

Use statistical terms carefully and correctly on the AP® Statistics exam. Don’t say “mean” if you really mean “median.” Range is a single number; so are Q_1 , Q_3 , and *IQR*. Avoid poor use of language, like “the outlier *skews* the mean” or “the median is in the middle of the *IQR*.” Skewed is a shape and the *IQR* is a single number, not a region. If you misuse a term, expect to lose some credit.

Here’s an activity that gives you a chance to put into practice what you have learned in this section.

ACTIVITY**Team challenge: Did Mr. Starnes stack his class?**

In this activity, you will work in a team of three or four students to resolve a dispute.

Mr. Starnes teaches AP® Statistics, but he also does the class scheduling for the high school. There are two AP® Statistics classes—one taught by Mr. Starnes and one taught by Ms. McGrail. The two teachers give the same first test to their classes and grade the test together. Mr. Starnes’s students earned an average score that was 8 points higher than the average for Ms. McGrail’s class. Ms. McGrail wonders whether Mr. Starnes might have “adjusted” the class rosters from the computer scheduling program. In other words, she thinks he might have “stacked” his class. He denies this, of course.

To help resolve the dispute, the teachers collect data on the cumulative grade point averages and SAT Math scores of their students. Mr. Starnes provides the GPA data from his computer. The students report their SAT Math scores. The following table shows the data for each student in the two classes.

Did Mr. Starnes stack his class? Give appropriate graphical and numerical evidence to support your conclusion. Be prepared to defend your answer.

ACTIVITY OVERVIEW

To prepare for using this activity, watch the overview video by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher’s Resources on the book’s digital platform.

Time: 30 minutes

Materials: A calculator or computer for each group of students

Teaching Advice: Students should use calculators, computer software, or an Extra Applet (highschool.bfwpub.com/updatedtps6e)

to analyze the data. At the end of the activity, pose the following question: If students were randomly assigned to the two teachers, how likely is it that there would be a difference in average test scores of 8 points? If the difference is so large that it is unlikely to happen purely by chance, we have convincing evidence that Mr. Starnes stacked the classes. In Chapter 11, this probability will be the *P*-value for a two-sample *t* test for the difference of means.

Student	Teacher	GPA	SAT-M
1	Starnes	2.900	670
2	Starnes	2.860	520
3	Starnes	2.600	570
4	Starnes	3.600	710
5	Starnes	3.200	600
6	Starnes	2.700	590
7	Starnes	3.100	640
8	Starnes	3.085	570
9	Starnes	3.750	710
10	Starnes	3.400	630
11	Starnes	3.338	630
12	Starnes	3.560	670
13	Starnes	3.800	650
14	Starnes	3.200	660
15	Starnes	3.100	510

Student	Teacher	GPA	SAT-M
16	McGrail	2.900	620
17	McGrail	3.300	590
18	McGrail	3.980	650
19	McGrail	2.900	600
20	McGrail	3.200	620
21	McGrail	3.500	680
22	McGrail	2.800	500
23	McGrail	2.900	502.5
24	McGrail	3.950	640
25	McGrail	3.100	630
26	McGrail	2.850	580
27	McGrail	2.900	590
28	McGrail	3.245	600
29	McGrail	3.000	600
30	McGrail	3.000	620
31	McGrail	2.800	580
32	McGrail	2.900	600
33	McGrail	3.200	600

You can use technology to make boxplots, as the following Technology Corner illustrates.

4. Technology Corner

MAKING BOXPLOTS

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

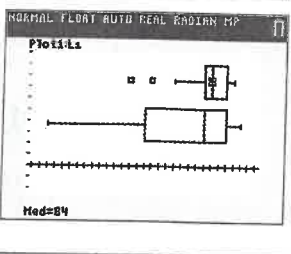
The TI-83/84 can plot up to three boxplots in the same viewing window. Let's use the calculator to make parallel boxplots of the overall rating data for Apple and Samsung tablets.

1. Enter the ratings for Apple tablets in list L1 and for Samsung in list L2.
2. Set up two statistics plots: Plot1 to show a boxplot of the Apple data in list L1 and Plot2 to show a boxplot of the Samsung data in list L2. The setup for Plot1 is shown. When you define Plot2, be sure to change L1 to L2.

Note: The calculator offers two types of boxplots: one that shows outliers and one that doesn't. We'll always use the type that identifies outliers.

3. Use the calculator's Zoom feature to display the parallel boxplots. Then Trace to view the five-number summary.

- Press **ZOOM** and select ZoomStat.
- Press **TRACE**.



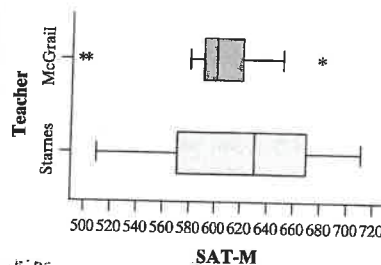
classes, the *IQR* is much larger for students in Mr. Starnes's class, mostly because the value of Q_3 is much higher for Mr. Starnes's class. Because the median and third quartile of the distribution of GPAs in Mr. Starnes's class are much higher than in Ms. McGrail's class, there is evidence that he stacked his class.

II. The SAT Math Data

Here are boxplots and summary statistics comparing the SAT Math data for Starnes and McGrail:

Variable	N	Mean	Median	StDev
Starnes SAT-M	15	622.0	630.0	61.3
McGrail SAT-M	18	600.1	600.0	44.1

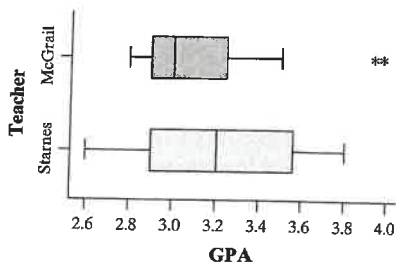
Variable	Min	Max	Q_1	Q_3
Starnes SAT-M	510.0	710.0	570.0	670.0
McGrail SAT-M	500.0	680.0	590.0	620.0



McGrail's distribution is somewhat symmetric, though there are two low outliers and one high outlier. Starnes's distribution of SAT Math scores is slightly skewed to the left. The median SAT Math score is higher for Starnes's class. The SAT Math scores vary more in Starnes's class, mostly due to the larger values for Q_3 and the maximum in Starnes's distribution. Again, because the median, Q_3 , and maximum are all higher for students in Mr. Starnes's class, it appears that Mr. Starnes stacked his class.

Teaching Tip: Using Technology

When making calculator boxplots, always have your students use the boxplot option that marks the outliers. The TI-Nspire automatically graphs boxplots this way.



The GPA distribution in Mr. Starnes's class is roughly symmetric, whereas the distribution for Ms. McGrail's class is right-skewed with two high outliers. The median GPA is higher in Mr. Starnes's class. Even though the range of GPAs is about the same for students in both

OVERVIEW (continued)

Answers:

1. The GPA Data

Here are boxplots and summary statistics comparing the GPA data for McGrail and Starnes:

Variable	N	Mean	Median	StDev
Starnes GPA	15	3.2129	3.2000	0.3642
McGrail GPA	18	3.1347	3.0000	0.3580

Variable	Min	Max	Q_1	Q_3
Starnes GPA	2.6000	3.8000	2.9000	3.5600
McGrail GPA	2.8000	3.9800	2.9000	3.2450

TRM Section 1.3 Quiz

There is one quiz available for this section in the Teacher's Resource Materials. Click on the link in the TE-Book, open in the TRFD, or download from the Teacher's Resources on the book's digital platform. You can also create your own quiz using the ExamView® Assessment Suite that is part of the TPS6 program. Questions are coded by Learning Target to make it easy to build parallel quizzes.

Section 1.3 Summary

- A numerical summary of a distribution should include measures of center and variability.
- The mean and the median describe the center of a distribution in different ways. The mean is the average of the observations. In symbols, the sample mean $\bar{x} = \frac{\sum x_i}{n}$. The median is the midpoint of the distribution, the number such that about half the observations are smaller and half are larger.
- A statistic is a number that describes a sample. A parameter is a number that describes a population. We often use statistics (like the sample mean $\bar{x}$) to estimate parameters (like the population mean μ).
- The simplest measure of variability for a distribution of quantitative data is the range, which is the distance from the maximum value to the minimum value.
- When you use the mean to describe the center of a distribution, use the standard deviation to describe the distribution's variability. The standard deviation gives the typical distance of the values in a distribution from the mean. In symbols, the sample standard deviation $s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$. The standard deviation s_x is 0 when there is no variability and gets larger as variability from the mean increases.
- The variance is the average of the squared deviations from the mean. The sample variance s_x^2 is the square of the sample standard deviation.
- When you use the median to describe the center of a distribution, use the interquartile range (IQR) to describe the distribution's variability. The first quartile Q_1 has about one-fourth of the observations below it, and the third quartile Q_3 has about three-fourths of the observations below it. The interquartile range measures variability in the middle half of the distribution and is found using $IQR = Q_3 - Q_1$.
- The median is a resistant measure of center because it is relatively unaffected by extreme observations. The mean is not resistant. Among measures of variability, the IQR is resistant, but the standard deviation and range are not.
- According to the $1.5 \times IQR$ rule, an observation is an outlier if it is less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$.
- The five-number summary of a distribution consists of the minimum, Q_1 , the median, Q_3 , and the maximum. A boxplot displays the five-number summary, marking outliers with a special symbol. The box shows the variability in the middle half of the distribution. Whiskers extend from the box to the smallest and the largest observations that are not outliers. Boxplots are especially useful for comparing distributions.

1.3 Technology Corners

Ti-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/updatedtps6e.

3. Computing numerical summaries

Page 66

4. Making boxplots

Page 73

57. Quiz g
period 3

86

87

(a) Calcula

(b) Suppos

quiz, ap

mean.

58. Pulse e

(in beat

71 10

74 10

(a) Calcula

(b) The sty

Find th

What p

59. Quiz g

(a) Find th

(b) Find th

unexcu

not hav

60. Pulse e

(a) Find th

(b) Find th

the me

rate do

61. Electig

United

majorit

to win

Here is

in 2016

Colum

(a) Find th

TRM Fu
Exercise

The full sc
the link in
download
the book's

Answers

1.87 (a) T
is 85. (b)
mean wor
of nonresi
sensitive

Section 1.3 Exercises

87. Quiz grades Joey's first 14 quiz grades in a marking period were as follows:

86	84	91	75	78	80	74
87	76	96	82	90	98	93

- (a) Calculate the mean. Show your work.
 (b) Suppose Joey has an unexcused absence for the 15th quiz, and he receives a score of 0. Recalculate the mean. What property of the mean does this illustrate?

88. Pulse rates Here are data on the resting pulse rates (in beats per minute) of 19 middle school students:

71	104	76	88	78	71	68	86	70	90
74	76	69	68	88	96	68	82	120	

- (a) Calculate the mean. Show your work.
 (b) The student with a 120 pulse rate has a medical issue. Find the mean pulse rate for the other 18 students. What property of the mean does this illustrate?

89. Quiz grades Refer to Exercise 87.

- (a) Find the median of Joey's first 14 quiz grades.
 (b) Find the median of Joey's quiz grades after his unexcused absence. Explain why the 0 quiz grade does not have much effect on the median.

90. Pulse rates Refer to Exercise 88.

- (a) Find the median pulse rate for all 19 students.
 (b) Find the median pulse rate excluding the student with the medical issue. Explain why this student's 120 pulse rate does not have much effect on the median.

91. Electing the president To become president of the United States, a candidate does not have to receive a majority of the popular vote. The candidate does have to win a majority of the 538 Electoral College votes. Here is a stemplot of the number of electoral votes in 2016 for each of the 50 states and the District of Columbia:

0	3333333344444
0	55566666677788999
1	00001111234
1	5668
2	00
2	99
3	
3	8
4	
4	
5	
5	5

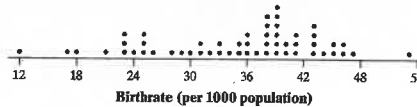
Key: 1|5 is a state with 15 electoral votes.

- (a) Find the median.

- (b) Without doing any calculations, explain how the mean and median compare.

- (c) Is the value you found in part (a) a statistic or a parameter? Justify your answer.

92. Birthrates in Africa One of the important factors in determining population growth rates is the birthrate per 1000 individuals in a population. The dotplot shows the birthrates per 1000 individuals (rounded to the nearest whole number) for all 54 African nations.



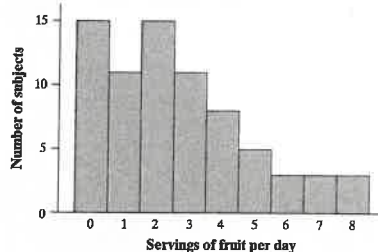
- (a) Find the median.
 (b) Without doing any calculations, explain how the mean and median compare.
 (c) Is the value you found in part (a) a statistic or a parameter? Justify your answer.

93. House prices The mean and median selling prices of existing single-family homes sold in September 2016 were \$276,200 and \$234,200.⁴⁰ Which of these numbers is the mean and which is the median? Explain your reasoning.

94. Mean salary? Last year a small accounting firm paid each of its five clerks \$32,000, two junior accountants \$60,000 each, and the firm's owner \$280,000.

- (a) What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary?
 (b) Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

95. Do adolescent girls eat fruit? We all know that fruit is good for us. Here is a histogram of the number of servings of fruit per day claimed by 74 seventeen-year-old girls in a study in Pennsylvania.⁴¹



- 1.90 (a) The median pulse rate for all 19 students is 76 beats per minute. (b) 76 beats per minute; notice that the median did not change. This shows that the median is resistant to outliers.

- 1.91 (a) 8 electoral votes (b) Since the distribution of number of electoral votes is skewed to the right, the mean of this distribution is greater than its median. (c) The median, calculated in part (a), is a parameter because it is a number that describes the population of all 50 states and the District of Columbia.

- 1.92 (a) The median is 37.5 births per 1000 individuals in the population.

- (b) Since the distribution of number of births per 1000 individuals in the population is slightly skewed to the left, the mean of this distribution is less than its median. (c) The median, calculated in part (a), is a parameter because it is a number that describes the population of all 54 African nations.

- 1.93 The mean house price is \$276,200 and the median is \$234,200. The distribution of house prices is likely to be quite skewed to the right because of a few very expensive homes, some of which may be outliers. When a distribution is skewed to the right, the mean is bigger than the median.

- 1.94 (a) The mean salary is \$70,000. Seven of the eight employees (everyone but the owner) earned less than the mean. The median salary is \$32,000. (b) An unethical recruiter would report the mean salary as the "typical" or "average" salary. However, the median is a more accurate depiction of a "typical" employee's earnings, because it is not influenced by the outlier (\$280,000).

Full Solutions to Section 1.3 Exercises

The full solutions can be found by clicking on the link in the TE-Book, opening the TRFD, or downloading from the Teacher's Resources on the book's digital platform.

Answers to Section 1.3 Exercises

- 1.87 (a) The mean of Joey's first 14 quiz scores is 85. (b) Including a 15th quiz score of 0, Joey's mean would be 79.3. This illustrates the property of nonresistance. The mean is not resistant. It is sensitive to extreme values.

- 1.88 (a) The mean of the resting pulse rates for the 19 middle school students is 81.21 beats per minute. (b) Deleting the student from the list with a resting pulse rate of 120 beats per minute, the new mean is 79.06 beats per minute. This illustrates the property of nonresistance. The mean is not resistant. It is sensitive to extreme values.

- 1.89 (a) 85 (b) Since there are now 15 quiz scores, the median is 84. Notice that the median did not change much. This shows that the median is resistant to outliers.

1.95 (a) The median is 2 servings of fruit per day. **(b)** $\bar{x} = 194/74 = 2.62$ servings of fruit per day

1.96 (a) The 50th percentile corresponds to a word length of 4. **(b)** The mean is $413/100 = 4.13$. Alternatively, we can estimate the mean by finding the balance point of the distribution, which appears to be slightly larger than 4.

1.97 (a) Range = Max - Min = $98 - 74 = 24$; the range of Joey's quiz grades after his unexcused absence is $98 - 0 = 98$. **(b)** The range may not be the best way to describe variability for a distribution of quantitative data because the range can be heavily affected by outliers.

1.98 (a) Range = 52 beats per minute; the range of the resting pulse rates excluding the student with the medical issue is $104 - 68 = 36$ beats per minute. **(b)** The range may not be the best way to describe variability for a distribution of quantitative data because the range can be heavily affected by outliers.

1.99 The standard deviation is 2.52 cm. The foot lengths of these 14-year-olds from the U.K. typically vary by about 2.52 cm from the mean of 24 cm.

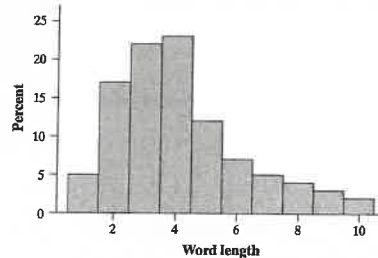
1.100 The standard deviation is 1.673 hours. The amount of sleep that these 6 students got last night typically varies by about 1.673 hours from the mean of 8 hours.

1.101 (a) The size of these 18 files typically varies by about 1.9 megabytes from the mean of 3.2 megabytes. **(b)** If the music file that takes up 7.5 megabytes of storage space is replaced with another version of the file that takes up only 4 megabytes, the mean would decrease slightly. The standard deviation would decrease as well because a file of size 4 megabytes will be closer to the new mean than the file of 7.5 megabytes was to the former mean.

1.102 (a) The fat content of these 12 hamburgers typically varies by about 9.06 grams from the mean of 22.83 grams. **(b)** If the restaurant replaces the burger that has 22 grams of fat with a burger that has 35 grams of fat, the mean would increase slightly. The standard deviation would increase as well because a hamburger with 35 grams of fat will be farther from the new mean than the hamburger of 22 grams was from the former mean.

1.103 Variable B has a smaller standard deviation because more of the observations have values closer to the mean than in Variable A's distribution.

- (a) Find the median number of servings of fruit per day from the histogram. Explain your method clearly.
- (b) Calculate the mean of the distribution. Show your work.
96. **Shakespeare** The histogram shows the distribution of lengths of words used in Shakespeare's plays.⁴²



- (a) Find the median word length in Shakespeare's plays from the histogram. Explain your method clearly.
- (b) Calculate the mean of the distribution. Show your work.
97. **Quiz grades** Refer to Exercise 87.
- (a) Find the range of Joey's first 14 quiz grades and the range of Joey's quiz grades after his unexcused absence.
- (b) Explain what part (a) suggests about using the range as a measure of variability for a distribution of quantitative data.

98. **Pulse rates** Refer to Exercise 88.

- (a) Find the range of the pulse rates for all 19 students and the range of the pulse rates excluding the student with the medical issue.
- (b) Explain what part (a) suggests about using the range as a measure of variability for a distribution of quantitative data.

99. **Foot lengths** Here are the foot lengths (in centimeters) for a random sample of seven 14-year-olds from the United Kingdom:

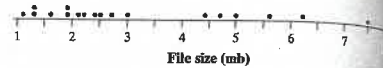
25	22	20	25	24	24	28
----	----	----	----	----	----	----

Calculate the sample standard deviation. Interpret this value.

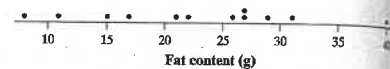
100. **Well rested?** A random sample of 6 students in a first-period statistics class was asked how much

sleep (to the nearest hour) they got last night. Their responses were 6, 7, 7, 8, 10, and 10. Calculate the sample standard deviation. Interpret this value.

101. **File sizes** How much storage space does your music use? Here is a dotplot of the file sizes (to the nearest tenth of a megabyte) for 18 randomly selected files on Nathaniel's mp3 player:

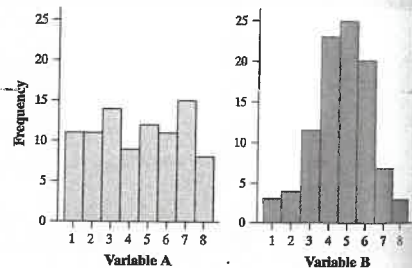


- (a) The distribution of file size has a mean of $\bar{x} = 3.2$ megabytes and a standard deviation of $s_x = 1.9$ megabytes. Interpret the standard deviation.
- (b) Suppose the music file that takes up 7.5 megabytes of storage space is replaced with another version of the file that only takes up 4 megabytes. How would this affect the mean and the standard deviation? Justify your answer.
102. **Healthy fast food?** Here is a dotplot of the amount of fat (to the nearest gram) in 12 different hamburgers served at a fast-food restaurant.



- (a) The distribution of fat content has a mean of $\bar{x} = 22.83$ grams and a standard deviation of $s_x = 9.06$ grams. Interpret the standard deviation.
- (b) Suppose the restaurant replaces the burger that has 22 grams of fat with a new burger that has 35 grams of fat. How would this affect the mean and the standard deviation? Justify your answer.

103. **Comparing SD** Which of the following distributions has a smaller standard deviation? Justify your answer.



That is, the typical distance from the mean is smaller for Variable B than for Variable A.

1.104 Machine B has a larger standard deviation because more of the observations have values farther from the mean than in Machine A's distribution.

1.105 $Q_1 = 1.9$ megabytes; $Q_3 = 4.7$; $IQR = Q_3 - Q_1 = 4.7 - 1.9 = 2.8$ megabytes

1.106 $Q_1 = 16$ grams; $Q_3 = 28$ grams; $IQR = Q_3 - Q_1 = 28 - 16 = 12$ grams

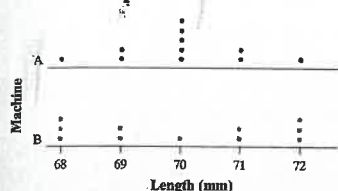
1.107 An outlier is any value below $Q_1 - 1.5(IQR) = 1.9 - 1.5(2.8) = -2.3$ megabytes or above $Q_3 + 1.5(IQR) = 4.7 + 1.5(2.8) = 8.9$ megabytes. No files in the data set have less than -2.3 megabytes or more than 8.9 megabytes; there are no outliers.

1.108 An outlier is any value below $Q_1 - 1.5(IQR) = 16 - 1.5(12) = -2$ grams of fat or above $Q_3 + 1.5(IQR) = 28 + 1.5(12) = 46$ grams of fat. No hamburgers in the data set have fewer than -2 grams of fat or more than 46 grams of fat, so there are no outliers in the distribution.

1.109 (a) The distribution is skewed to the right because the mean is much larger than the median. Also, Q_3 is much farther from the median than Q_1 . **(b)** The amount of money spent typically varies by about \$21.70 from the mean of \$34.70. **(c)** $Q_1 = 19.27$; $Q_3 = 45.40$; so the IQR is $45.40 - 19.27 = 26.13$. Any points below $19.27 - 1.5(26.13) = -19.925$ or above $45.40 + 1.5(26.13) = 84.595$ are outliers. Because the maximum of 93.34 is greater than 84.595, there is at least one outlier.

1.110 (a) T is more likely and median (19.1 and median for apart (41.3 the range standard (and is not middle 50 spreads, if distributio distributio standard (Any points above 50 Because (least one

104. **Comparing SD** The parallel dotplots show the lengths (in millimeters) of a sample of 11 nails produced by each of two machines. Which distribution has the larger standard deviation? Justify your answer.



105. **File sizes** Refer to Exercise 101. Find the interquartile range of the file size distribution shown in the dotplot.

106. **Healthy fast food?** Refer to Exercise 102. Find the interquartile range of the fat content distribution shown in the dotplot.

107. **File sizes** Refer to Exercises 101 and 105. Identify any outliers in the distribution. Show your work.

108. **Healthy fast food?** Refer to Exercises 102 and 106. Identify any outliers in the distribution. Show your work.

109. **Shopping spree** The figure displays computer output for data on the amount spent by 50 grocery shoppers.

	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
Amount spent	34.70	21.70	3.11	19.27	27.86	45.40	93.34

- (a) What would you guess is the shape of the distribution based only on the computer output? Explain.
 (b) Interpret the value of the standard deviation.
 (c) Are there any outliers? Justify your answer.

110. **C-sections** A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a year by samples of male and female doctors. Here are summary statistics for the two distributions:

	$\bar{x}$	s_x	Min	Q_1	Med	Q_3	Max
Male doctors	41.333	20.607	20	27	34	50	86
Female doctors	19.1	10.126	5	10	18.5	29	33

- (a) Based on the computer output, which distribution would you guess has a more symmetrical shape? Explain your answer.

- (b) Explain how the IQRs of these two distributions can be so similar even though the standard deviations are quite different.
 (c) Does either distribution have any outliers? Justify your answer.

111. **Don't call me** According to a study by Nielsen Mobile, "Teenagers ages 13 to 17 are by far the most prolific texters, sending 1742 messages a month." Mr. Williams, a high school statistics teacher, was skeptical about the claims in the article. So he collected data from his first-period statistics class on the number of text messages they had sent in the past 24 hours. Here are the data:

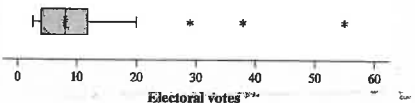
0	7	1	29	25	8	5	1	25	98	9	0	26
8	118	72	0	92	52	14	3	3	44	5	42	

- (a) Make a boxplot of these data.
 (b) Use the boxplot you created in part (a) to explain how these data seem to contradict the claim in the article.
112. **Acing the first test** Here are the scores of Mrs. Liao's students on their first statistics test:

93	93	87.5	91	94.5	72	96	95	93.5	93.5	73
82	45	88	80	86	85.5	87.5	81	78	86	89
92	91	98	85	82.5	88	94.5	43			

- (a) Make a boxplot of these data.
 (b) Use the boxplot you created in part (a) to describe how the students did on Mrs. Liao's first test.

113. **Electing the president** Refer to Exercise 91. Here are a boxplot and some numerical summaries of the electoral vote data:

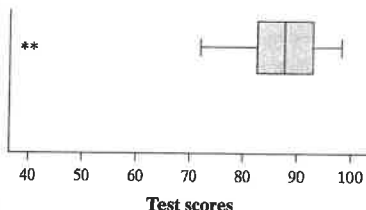


Variable	N	Mean	SD	Min	Q_1	Median	Q_3	Max
Electoral votes	51	10.55	9.69	3	4	8	12	55

- (a) Explain why the median and IQR would be a better choice for summarizing the center and variability of the distribution of electoral votes than the mean and standard deviation.
 (b) Identify an aspect of the distribution that the stemplot in Exercise 91 reveals that the boxplot does not.

- (b) The article claims that teens send 1742 texts a month, or about 58 texts a day (assuming a 30-day month). Nearly all (21 of 25) sent fewer than 58 texts per day, which seems to contradict the claim in the article.

- 1.112 (a) The median is 87.75, $Q_1 = 82$; $Q_3 = 93$. The IQR is $93 - 82 = 11$. Any observation above $Q_3 + 1.5/IQR = 93 + 1.5(11) = 109.5$ or below $Q_1 - 1.5/IQR = 82 - 1.5(11) = 65.5$ is considered an outlier. Thus, the scores 43 and 45 are outliers. The boxplot is shown below.



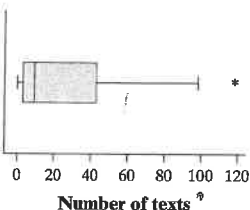
- (b) Most students did quite well. In fact, more than 75% of the class scored higher than 80. Only two of the students did very poorly.

- 1.113 (a) The median and IQR would be a better choice for summarizing the center and variability of the distribution of electoral votes than the mean and standard deviation because the boxplot reveals that there are three outliers in the data set. The mean and standard deviation are not resistant measures of center and variability, so their values are sensitive to these extreme values. (b) The stemplot reveals that the distribution has a single peak, which cannot be discerned from the boxplot. Also, the stemplot reveals that there are actually four upper outliers rather than three. The value of 29, which is an outlier, gives the number of electoral votes for two states. In the boxplot, this appears as one asterisk. However, there are two states that have that many electoral votes, not one.

- 1.110 (a) The distribution for the female doctors is more likely to be symmetric since the mean and median are relatively close together (19.1 and 18.5, respectively). The mean and median for the male doctors are quite far apart (41.333 and 34). (b) The IQR measures the range of the middle 50% of the data. The standard deviation, uses every observation and is not resistant to outliers. So, while the middle 50% of the data sets may have similar spreads, if the lower and upper 25% of one distribution are more variable than the other distribution, the distribution will have a larger standard deviation. (c) Male distribution: Any points below $27 - 1.5(23) = -7.5$ or above $50 + 1.5(23) = 84.5$ are outliers. Because 86 is greater than 84.5, there is at least one upper outlier. There are no lower

- outliers. Female distribution: Any points below $10 - 1.5(19) = -18.5$ or above $29 + 1.5(19) = 57.5$ are outliers. This distribution does not have any outliers.

- 1.111 (a) The median is 9, the first quartile is 3, and the third quartile is 43. The IQR is $43 - 3 = 40$. An outlier would be any value below $3 - 1.5(40) = -57$ or above $43 + 1.5(40) = 103$. This means that the value of 118 is an outlier. The boxplot is shown below.

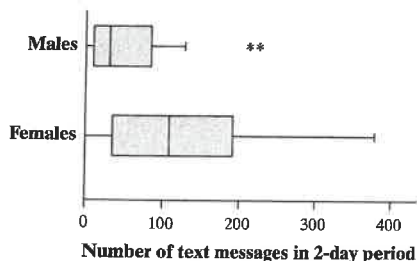


1.114 (a) The median and *IQR* would be a better choice for summarizing the center and variability of the distribution of birthrates in African countries because the boxplot is left-skewed. We can confirm the left skew of this distribution by noticing that the distance from median to minimum is 23.5, while the distance from median to maximum is 15.5. Also, the mean is less than the median. **(b)** The dotplot reveals that there is a single peak at 38 births per 1000 individuals in the population. This cannot be discerned from the boxplot.

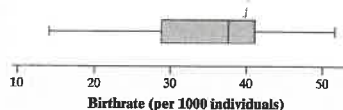
1.115 Shape: The distribution of energy cost (in dollars) for top freezers looks roughly symmetric. The distribution of energy cost (in dollars) for side freezers looks roughly symmetric. The distribution of energy cost (in dollars) for bottom freezers looks skewed to the right. **Outliers:** There are no outliers for the top or side freezers. There are at least two bottom freezers with unusually high energy costs (over \$140 per year). **Center:** The typical energy cost for the side freezers (median $\approx$ \$75) is greater than the typical cost for the bottom freezers (median $\approx$ \$69), which is greater than the typical cost for the top freezers (median $\approx$ \$56). **Variability:** There is much more variability in the energy costs for bottom freezers (*IQR* $\approx$ \$20), than for side freezers (*IQR* $\approx$ \$12), than for top freezers (*IQR* $\approx$ \$8).

1.116 Shape: The distributions of annual household income for Maine and Massachusetts are both skewed to the right, while the distribution for Connecticut is fairly symmetric. **Outliers:** There is one outlier in the distribution for Maine at $\approx$ \$160,000, and one outlier in the distribution for Massachusetts at $\approx$ \$430,000. There are no outliers in the Connecticut distribution. **Center:** The center of the distribution is lowest for Maine (median $\approx$ \$43,000), followed by Connecticut (median $\approx$ \$60,000), and then Massachusetts (median $\approx$ \$95,000). **Variability:** The variability of the household income distribution is similar for Maine (*IQR* $\approx$ \$47,000) and Connecticut (*IQR* $\approx$ \$50,000), but much more variable for Massachusetts (*IQR* $\approx$ \$130,000).

1.117 (a)



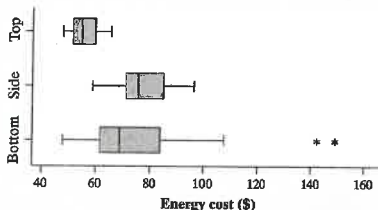
114. Birthrates in Africa Refer to Exercise 92. Here are a boxplot and some numerical summaries of the birthrate data:



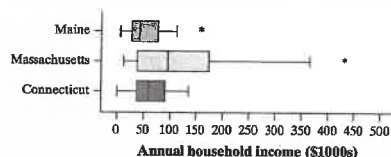
Variable	N	Mean	SD	Min	Q ₁	Median	Q ₃	Max
Birthrate	54	34.91	8.57	14.00	29.00	37.50	41.00	53.00

- (a) Explain why the median and *IQR* would be a better choice for summarizing the center and variability of the distribution of birthrates in African countries than the mean and standard deviation.
- (b) Identify an aspect of the distribution that the dotplot in Exercise 92 reveals that the boxplot does not.

115. Energetic refrigerators *Consumer Reports* magazine rated different types of refrigerators, including those with bottom freezers, those with top freezers, and those with side freezers. One of the variables they measured was annual energy cost (in dollars). The following boxplots show the energy cost distributions for each of these types. Compare the energy cost distributions for the three types of refrigerators.



116. Income in New England The following boxplots show the total income of 40 randomly chosen households each from Connecticut, Maine, and Massachusetts, based on U.S. Census data from the American Community Survey. Compare the distributions of annual incomes in the three states.



117. Who texts more? For their final project, a group of AP® Statistics students wanted to compare the texting habits of males and females. They asked a random

sample of students from their school to record the number of text messages sent and received over a two-day period. Here are their data:

Males	127	44	28	83	0	6	78	6
	5	213	73	20	214	28	11	
Females	112	203	102	54	379	305	179	24
	127	65	41	27	298	6	130	0

- (a) Make parallel boxplots of the data.
- (b) Use your calculator to compute separate numerical summaries for the males and for the females. Are these values statistics or parameters? Explain your answer.
- (c) Do these data suggest that males and females at the school differ in their texting habits? Use the results from parts (a) and (b) to support your answer.

118. SSHA scores Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for a random sample of 18 first-year college women:

154	109	137	115	152	140	154	178	101
103	126	126	137	165	165	129	200	148

Here are the SSHA scores for a random sample of 20 first-year college men:

108	140	114	91	180	115	126
92	169	146	109	132	75	88
113	151	70	115	187	104	

Note that high scores indicate good study habits and attitudes toward learning.

- (a) Make parallel boxplots of the data.
- (b) Use your calculator to compute separate numerical summaries for the women and for the men. Are these values statistics or parameters? Explain your answer.
- (c) Do these data support the belief that men and women differ in their study habits and attitudes toward learning? Use your results from parts (a) and (b) to support your answer.

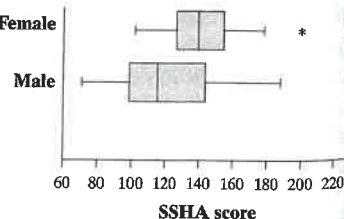
119. Income and education level Each March, the Bureau of Labor Statistics compiles an Annual Demographic Supplement to its monthly Current Population Survey.⁴³ Data on about 71,067 individuals between the ages of 25 and 64 who were employed full-time were collected in one of these surveys. The parallel boxplots compare the distributions of income for people with five levels of education. This figure is a variation of the boxplot idea: because large data sets often contain very extreme observations, we omitted the individuals in each category with the top 5% and bottom 5% of incomes. Also, the whiskers are drawn all the way to the maximum and minimum values of the remaining data for each distribution.

(b) Numerical summaries are given below.

	n	$\bar{X}$	S_x	Min	Q ₁	Median	Q ₃	Max
Male	15	62.4	71.4	0	6	28	83	214
Female	16	128.3	116.0	0	34	107	191	379

These values are statistics because they are numbers that describe the sample of students. **(c)** The data give very strong evidence that male and female texting habits differ considerably at the school. A typical female sends and receives about 79 more text messages in 2 days than a typical male. The males as a group are also much more consistent in their texting frequency than the females.

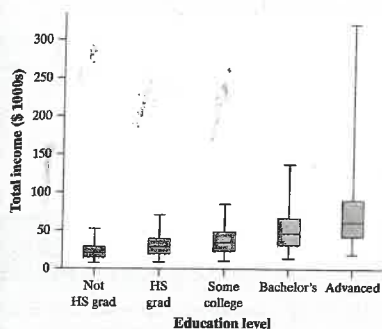
1.118 (a) Female



(b)

Variable	n	$\bar{X}$	S_x	Min	Q ₁	Median	Q ₃
Women	18	141.06	26.44	101.00	126.00	138.50	154.00
Men	20	121.25	32.85	70.00	98.00	114.50	143.00

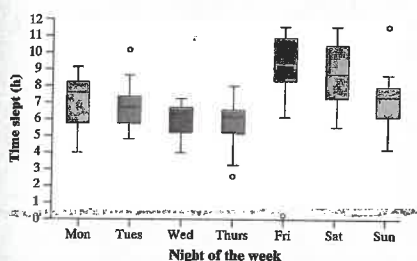
(c) The data give very strong evidence that first-year male and female college students' study habits and attitudes differ considerably. A typical first-year female earns, on average, about 20 more points on the SSHA than a typical male first-year student.



Use the graph to help answer the following questions.

- What shape do the distributions of income have?
- Explain how you know that there are outliers in the group that earned an advanced degree.
- How does the typical income change as the highest education level reached increases? Why does this make sense?
- Describe how the variability in income changes as the highest education level reached increases.

120. **Sleepless nights** Researchers recorded data on the amount of sleep reported each night during a week by a random sample of 20 high school students. Here are parallel boxplots comparing the distribution of time slept on all 7 nights of the study.⁴⁴



Use the graph to help answer the following questions.

- Which distributions have a clear left-skewed shape?
- Which outlier stands out the most, and why?

- How does the typical amount of sleep that the students got compare on these seven nights?
- On which night was there the most variation in how long the students slept? Justify your answer.

121. **SD contest** This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

- Choose four numbers that have the smallest possible standard deviation.
- Choose four numbers that have the largest possible standard deviation.
- Is more than one choice possible in either part (a) or (b)? Explain.

122. **What do they measure?** For each of the following summary statistics, decide (i) whether it could be used to measure center or variability and (ii) whether it is resistant.

- $\frac{Q_1 + Q_3}{2}$
- $\frac{\text{Max} - \text{Min}}{2}$

Multiple Choice: Select the best answer for Exercises 123–126.

123. If a distribution is skewed to the right with no outliers, which expression is correct?

- mean < median
- mean $\approx$ median
- mean = median
- mean > median
- We can't tell without examining the data.

124. The scores on a statistics test had a mean of 81 and a standard deviation of 9. One student was absent on the test day, and his score wasn't included in the calculation. If his score of 84 was added to the distribution of scores, what would happen to the mean and standard deviation?

- Mean will increase, and standard deviation will increase.
- Mean will increase, and standard deviation will decrease.
- Mean will increase, and standard deviation will stay the same.
- Mean will decrease, and standard deviation will increase.
- Mean will decrease, and standard deviation will decrease.

amounts. (b) The outlier of 0 hours on sleep on Friday night stands out the most. It is very unusual for a student to get 0 hours of sleep! (c) Students tended to get more sleep on Friday and Saturday nights than the other five nights of the week. (d) The most variation seems to exist in the number of hours that the students slept on Friday night. That evening, the number of hours the student slept varied from 0 to 12 hours, which is greater than every other night.

1.121 (a) One possible answer is 1, 1, 1, and 1. (b) 0, 0, 10, 10 (c) For part (a), any set of four identical numbers will have $s_x = 0$. For part (b), however, there is only 1 possible answer. We want the values to be as far from the mean as possible, so the squared deviations from the mean can be as big as possible. Our best choice is two values at each extreme, which makes all four squared deviations equal to 5^2 .

1.122 (a) This could be used to measure the center since we are averaging the 25th and 75th percentiles, effectively finding a middle point between these positions. It would be resistant to outliers, because any outliers would occur below Q_1 or above Q_3 . (b) This could be used as a measure of variability since it finds the distance between the smallest and largest values and then divides by 2, giving half of the range. This measure, however, would not be resistant to outliers. If outliers exist, they would, by definition, include the max, the min, or both.

1.123 d

1.124 b

1.119 (a) All five income distributions are skewed to the right. This tells me that within each level of education, there is much more variability in incomes for the upper 50% of the distribution than for the lower 50% of the distribution. (b) Even though the boxplots are not modified (i.e., do not show outliers), it is clear that the upper whisker for the advanced degree boxplot is much longer than 1.5 times its IQR, indicating that there is at least one upper outlier in the group that earned an advanced degree. (c) As education level rises, the median, quartiles, and extremes rise as well—that is, every value in

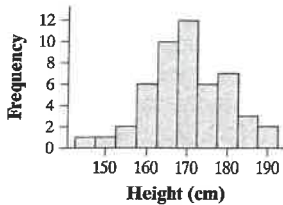
the 5-number summary gets larger. This makes sense because one would expect that individuals with more education tend to attain jobs that yield more income. (d) The variability in income increases as the education level increases. Both the width of the box (the IQR) and the distance from one extreme to the other increase as education levels increase.

1.120 (a) Monday, Wednesday, and Thursday nights' distributions have a clear left-skewed shape. This shape tells us that, on those nights, there is much more variability in the lower 50% of sleep amounts than the upper 50% of sleep

1.125 e

1.126 a

1.127 A histogram is given below.



This distribution is roughly symmetric, with a single peak at 170 cm. There do not appear to be any outliers. The center of the distribution of heights can be described by the mean of 169.88 cm or the median of 169.5 cm. The heights vary from 145.5 to 191 cm, so the range is 45.5 cm. The standard deviation of heights is 9.687 cm and the *IQR* is $177 - 163 = 14$ cm.

1.128 Women appear to be more likely to engage in behaviors that are indicative of good "habits of mind." They are especially more likely to revise papers to improve their writing (about 55% of females report this as opposed to about 37% of males). The difference is a little less for seeking feedback on their work. In that case, about 49% of the females did this as opposed to about 38% of males.

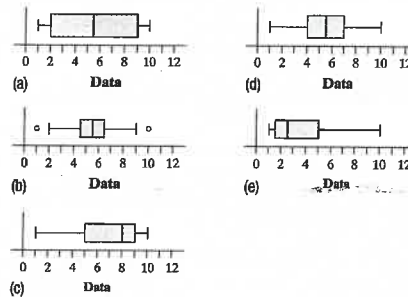
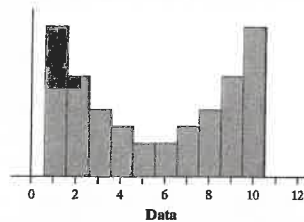
125. The stemplot shows the number of home runs hit by each of the 30 Major League Baseball teams in a single season. Home run totals above what value should be considered outliers?

09	15
10	3789
11	47
12	19
13	
14	89
15	34445
16	239
17	223
18	356
19	1
20	3
21	0
22	2

Key: 14|8 is a team with 148 home runs.

- (a) 173 (b) 210 (c) 222
(d) 229 (e) 257

126. Which of the following boxplots best matches the distribution shown in the histogram?



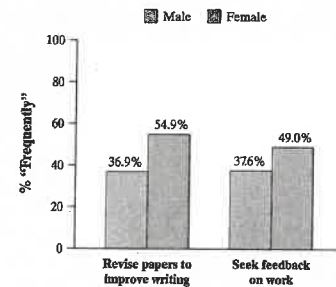
Recycle and Review

127. How tall are you? (1.2) We used Census At School's "Random Data Selector" to choose a sample of 50 Canadian students who completed a survey in a recent year. Here are the students' heights (in centimeters):

166.5	170.0	178.0	163.0	150.5	169.0	173.0	169.0	171.0	166.0
190.0	183.0	178.0	161.0	171.0	170.0	191.0	168.5	178.5	173.0
175.0	160.5	166.0	164.0	163.0	174.0	160.0	174.0	182.0	167.0
166.0	170.0	170.0	181.0	171.5	160.0	178.0	157.0	165.0	187.0
168.0	157.5	145.5	156.0	182.0	168.5	177.0	162.5	160.5	185.5

Make an appropriate graph to display these data. Describe the shape, center, and variability of the distribution. Are there any outliers?

128. Success in college (1.1) The Freshman Survey asked first-year college students about their "habits of mind"—specific behaviors that college faculty have identified as being important for student success. One question asked students, "How often in the past year did you revise your papers to improve your writing?" Another asked, "How often in the past year did you seek feedback on your academic work?" The figure is a bar graph comparing the percent of males and females who answered "frequently" to these two questions.⁴⁵



What does the graph reveal about the habits of mind of male and female college freshmen?

The free r
else r
Dire
ods)
of yo
ness
Usir
of 3
Amc
fema
(ma
T
betw

Ch

Introdu
In this l
that wil
tributio
ical var.
• Beg
on
• Sta
sun

✚ As
How t
Statis
This p
great /
resour