CHAPTER 3

Numerically Summarizing Data

Outline

DECISIONS

Suppose that you are in the market for a used car. To make an informed decision regarding your purchase, you decide to collect as much information as possible. What information is important in helping you make this decision? See the Decisions project on page 164.

Putting It All Together

When we look at a distribution of data, we should consider three characteristics of the distribution: its shape, its center, and its spread. In the last chapter, we discussed methods for organizing raw data into tables and graphs. These graphs (such as the histogram) allow us to identify the shape of the distribution. Recall that we describe the shape of a distribution as symmetric (in particular, bell shaped or uniform), skewed right, or skewed left.

The center and spread are numerical summaries of the data. The center of a data set is commonly called the average. There are many ways to describe the average value of a distribution. In addition, there are many ways to measure the spread of a distribution. The most appropriate measure of center and spread depends on the shape of the distribution.

Once these three characteristics of the distribution are known, we can analyze the data for interesting features, including unusual data values, called *outliers*.

3.1 Measures of Central Tendency

Preparing for This Section Before getting started, review the following:

- Quantitative data (Section 1.1, p. 8)
- Qualitative data (Section 1.1, p. 8)
- Population versus sample (Section 1.1, p. 4)
- Simple random sampling (Section 1.2, pp. 16–19)

Objectives

1 **Determine the arithmetic mean of a variable from raw data**

2 **Determine the median of a variable from raw data**

3 **Determine the mode of a variable from raw data**

4 **Use the mean and the median to help identify the shape of a distribution**

A measure of central tendency numerically describes the average or typical data value of a variable. We hear the word *average* in the news all the time:

- The average miles per gallon of gasoline of the 2006 Chevrolet Camaro in city driving is 19 miles.
- According to the U.S. Census Bureau, the national average commute time to work in 2005 was 24.3 minutes.
- According to the U.S. Census Bureau, the average household income in 2003 was $43,527.
- The average American woman is 5′4″ tall and weighs 142 pounds.

CAUTION

Whenever you hear the word *average*, be aware that the word may not always be referring to the mean. One average could be used to support one position, while another average could be used to support a different position.

In this chapter, we discuss three measures of central tendency: the *mean*, the *median*, and the *mode*. While other measures of central tendency exist, these three are the most widely used. When the word *average* is used in the media (newspapers, reporters, and so on) it usually refers to the mean. But beware! Some reporters use the term *average* to refer to the median or mode. As we shall see, these three measures of central tendency can give very different results!

Before we discuss measures of central tendency, we must consider whether or not we are computing a measure of central tendency that describes a population or one that describes a sample.

Definitions

A **parameter** is a descriptive measure of a population.

A **statistic** is a descriptive measure of a sample.

In Other Words

To help you remember the difference between a parameter and a statistic, think of the following:

$\underline{p}$ = $\underline{p}$arameter = $\underline{p}$opulation

$\underline{s}$ = $\underline{s}$tatistic = $\underline{s}$ample

For example, if we determine the average test score for *all* the students in a statistics class, our population, the average is a parameter. If we compute the average based on a simple random sample of five students, the average is a statistic.

1 Determine the Arithmetic Mean of a Variable from Raw Data

When used in everyday language, the word *average* often stands for the arithmetic mean. To compute the arithmetic mean of a set of data, the data must be quantitative.

Definitions

The **arithmetic mean** of a variable is computed by determining the sum of all the values of the variable in the data set, divided by the number of observations. The **population arithmetic mean**, μ (pronounced "mew"), is computed using all the individuals in a population. The population mean is a parameter. The **sample arithmetic mean**, $\bar{x}$ (pronounced "x-bar"), is computed using sample data. The sample mean is a statistic.

While other types of means exist (see Problems 51 and 52), the arithmetic mean is generally referred to as the **mean**. We will follow this practice for the remainder of the text.

In statistics, Greek letters are used to represent parameters, and Roman letters are used to represent statistics. Statisticians use mathematical expressions to describe the method for computing means.

Definitions

If $x_1, x_2, \ldots, x_N$ are the N observations of a variable from a population, then the population mean, μ, is

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N} \qquad \textbf{(1)}$$

If $x_1, x_2, \ldots, x_n$ are n observations of a variable from a sample, then the sample mean, $\bar{x}$, is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n} \qquad \textbf{(2)}$$

In Other Words

To find the mean of a set of data, add up all the observations and divide by the number of observations.

Note that N represents the size of the population, while n represents the size of the sample. The symbol Σ (the Greek letter capital sigma) tells us the terms are to be added. The subscript i is used to make the various values distinct and does not serve as a mathematical operation. For example, x_1 is the first data value, x_2 is the second, and so on.

Let's look at an example to help distinguish the population mean and sample mean.

EXAMPLE 1 Computing a Population Mean and a Sample Mean

Table 1

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Problem: The data in Table 1 represent the first exam score of 10 students enrolled in a section of Introductory Statistics.

(a) Compute the population mean.
(b) Find a simple random sample of size $n = 4$ students.
(c) Compute the sample mean of the sample obtained in part (b).

Approach

(a) To compute the population mean, we add up all the data values (test scores) and then divide by the number of individuals in the population.
(b) Recall from Section 1.2 that we can use either Table I in Appendix A, a calculator with a random-number generator, or computer software to obtain simple random samples. We will use a TI-84 Plus graphing calculator.
(c) The sample mean is found by adding the data values that correspond to the individuals selected in the sample and then dividing by $n = 4$, the sample size.

Solution

(a) We compute the population mean by adding the scores of all 10 students:

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + \cdots + x_{10}\\ &= 82 + 77 + 90 + 71 + 62 + 68 + 74 + 84 + 94 + 88\\ &= 790\end{aligned}$$

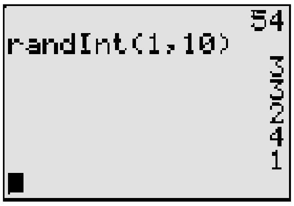
Divide this result by 10, the number of students in the class.

$$\mu = \frac{\sum x_i}{N} = \frac{790}{10} = 79$$

Although it was not necessary in this problem, we will agree to round the mean to one more decimal place than that in the raw data.

(b) To find a simple random sample of size $n = 4$ from a population whose size is $N = 10$, we will use the TI-84 Plus random-number generator with a seed of 54. (Recall that this gives the starting point that the calculator uses to generate the list of random numbers.) Figure 1 shows the students in the sample. Bilal (90), Ryanne (77), Pam (71), and Michelle (82) are in the sample.

Figure 1



```
                   54
randInt(1,10)
                    3
                    3
                    2
                    4
                    1
```

(c) We compute the sample mean by first adding the scores of the individuals in the sample.

$$\begin{aligned}\sum x_i &= x_1 + x_2 + x_3 + x_4 \\ &= 90 + 77 + 71 + 82 \\ &= 320\end{aligned}$$

Divide this result by 4, the number of individuals in the sample.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{320}{4} = 80$$

Now Work Problem 25.

In-Class Activity: Population Mean versus Sample Mean

Treat the students in the class as a population. All the students in the class should determine their pulse rates.

(a) Compute the population mean pulse rate.

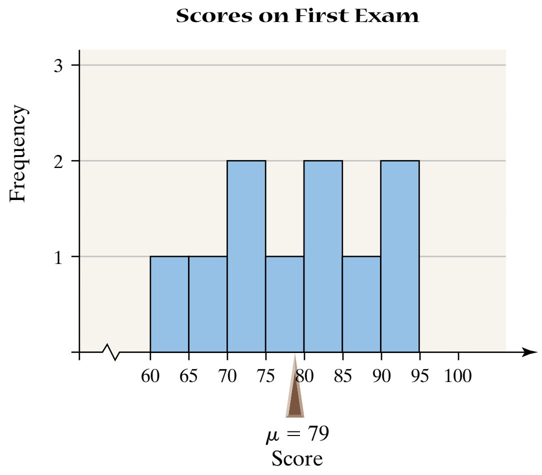
(b) Obtain a simple random sample of $n = 4$ students and compute the sample mean. Does the sample mean equal the population mean?

(c) Obtain a second simple random sample of $n = 4$ students and compute the sample mean. Does the sample mean equal the population mean?

(d) Are the sample means the same? Why?

It is helpful to think of the mean of a data set as the center of gravity. In other words, the mean is the value such that a histogram of the data is perfectly balanced, with equal weight on each side of the mean. Figure 2 shows a histogram of the data in Table 1 with the mean labeled. The histogram balances at $\mu = 79$.

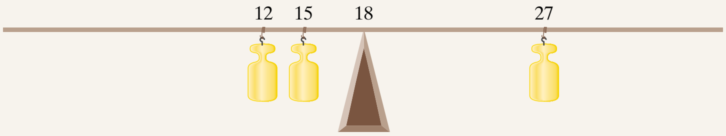
Figure 2



In-Class Activity: The Mean as the Center of Gravity

Find a yardstick, a fulcrum, and three objects of equal weight (maybe 1-kilogram weights from the physics department). Place the fulcrum at 18 inches so that the yardstick balances like a teeter-totter. Now place one weight on the yardstick at 12 inches, another at 15 inches, and the third at 27 inches. See Figure 3.

Figure 3



Does the yardstick balance? Now compute the mean of the location of the three weights. Compare this result with the location of the fulcrum. Conclude that the mean is the center of gravity of the data set.

2 Determine the Median of a Variable from Raw Data

A second measure of central tendency is the median. To compute the median of a set of data, the data must be quantitative.

Definition

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data are below the median and half the data are above the median. We use M to represent the median.

In Other Words

To help remember the idea behind the median, think of the median of a highway; it divides the highway in half.

To compute the median of a set of data, we use the following steps:

Steps in Computing the Median of a Data Set

Step 1**:** Arrange the data in ascending order.

Step 2**:** Determine the number of observations, n.

Step 3**:** Determine the observation in the middle of the data set.

- If the number of observations is odd, then the median is the data value that is exactly in the middle of the data set. That is, the median is the observation that lies in the $\left(\frac{n+1}{2}\right)$ position.
- If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the data values on either side of the observation that lies in the $\left(\frac{n+1}{2}\right)$ position.

EXAMPLE 2 Computing the Median of a Data Set with an Odd Number of Observations

Problem**:** The data in Table 2 represent the length (in seconds) of a random sample of songs released in the 1970s. Find the median length of the songs.

Approach**:** We will follow the steps listed above.

Solution

Step 1**:** Arrange the data in ascending order:

179, 201, 206, 208, 217, 222, 240, 257, 284

Table 2

Song Name	Length
“Sister Golden Hair”	201
“Black Water”	257
“Free Bird”	284
“The Hustle”	208
“Southern Nights”	179
“Stayin’ Alive”	222
“We Are Family”	217
“Heart of Glass”	206
“My Sharona”	240

***Step 2*:** There are $n = 9$ observations.

***Step 3*:** Since there are an odd number of observations, the median will be the observation exactly in the middle of the data set. The median, M, is 217 seconds (the $\frac{n+1}{2} = \frac{9+1}{2} = 5$th data value). We list the data in ascending order, with the median in blue.

$$179, 201, 206, 208, 217, 222, 240, 257, 284$$

Notice there are four observations to the left and four observations to the right of the median. We conclude that 50% of the songs are less than 217 minutes and 50% of the songs are more than 217 minutes.

EXAMPLE 3 Computing the Median of a Data Set with an Even Number of Observations

Problem: Find the median score of the data in Table 1 on page 108.

Approach: We will follow the steps given on page 110.

Solution

***Step 1*:** Arrange the data in ascending order:

$$62, 68, 71, 74, 77, 82, 84, 88, 90, 94$$

***Step 2*:** There are $n = 10$ observations.

***Step 3*:** Because there are $n = 10$ observations, the median will be the mean of the two middle observations. Because $\frac{n+1}{2} = \frac{10+1}{2} = 5.5$, the median is halfway between the fifth and sixth observations. We compute the median, M, by determining the mean of the fifth and sixth observations with the data written in ascending order. So the median is the mean of 77 and 82:

$$M = \frac{77 + 82}{2} = 79.5$$

Notice that there are five observations to the left and five observations to the right of the median, as follows:

$$62, 68, 71, 74, 77, \underset{\uparrow}{} 82, 84, 88, 90, 94$$

$$\boxed{M = 79.5}$$

We conclude that 50% of the students scored less than 79.5 and 50% of the students scored above 79.5.

Now compute the median of the data in Problem 19 by hand.

3 Determine the Mode of a Variable from Raw Data

A third measure of central tendency is the mode. The mode can be computed for either quantitative or qualitative data.

Definition

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

To compute the mode, tally the number of observations that occur for each data value. The data value that occurs most often is the mode. A set of data can have no mode, one mode, or more than one mode. If no observation occurs more than once, we say the data have **no mode**.

EXAMPLE 4 Finding the Mode of Quantitative Data

Problem: The following data represent the number of O-ring failures on the shuttle *Columbia* prior to its fatal flight for its seventeen flights:

0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3

Find the mode number of O-ring failures.

Approach: We tally the number of times we observe each data value. The data value with the highest frequency is the mode.

Solution: The mode is 0 because it occurs most frequently (eleven times).

EXAMPLE 5 Finding the Mode of Quantitative Data

Problem: Find the mode of the data listed in Table 1 on page 108.

Approach: Tally the number of times we observe each data value. The data value with the highest frequency is the mode. Although not necessary, it is helpful to find the mode of quantitative data by arranging the data in ascending order.

Solution: We arrange the data in ascending order:

62, 68, 71, 74, 77, 82, 84, 88, 90, 94

Since each data value occurs only once, there is no mode.

Now compute the mode of the data in Problem 19.

A data set can have more than one mode. For example, suppose the instructor added the scores of Pam and Sam incorrectly and they actually scored 77 and 88, respectively. The data set in Table 1 would now have two modes: 77 and 88. In this case, we say the data are **bimodal**. If a data set has three or more data values that occur with the highest frequency, the data set is **multimodal**. Typically, the mode is not reported for multimodal data because it is not representative of a central tendency or typical value.

We cannot determine the value of the mean or median of data that are qualitative. The only measure of central tendency that can be determined for qualitative data is the mode.

EXAMPLE 6 Determining the Mode of Qualitative Data

Problem: The data in Table 3 represent the location of injuries that required rehabilitation by a physical therapist. Determine the mode area of injury.

Table 3

Back	Back	Hand	Neck	Knee	Knee
Wrist	Back	Groin	Shoulder	Shoulder	Back
Elbow	Back	Back	Back	Back	Back
Back	Shoulder	Shoulder	Knee	Knee	Back
Hip	Knee	Hip	Hand	Back	Wrist

Source: Krystal Catton, student at Joliet Junior College

Approach: Determine the location of injury that occurs with the highest frequency.

Solution: The mode location of injury is the back, with 12 instances.

Now Work Problem 39.

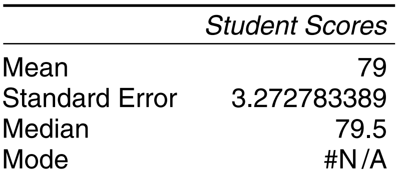
EXAMPLE 7 Finding the Mean, Median, and Mode Using Technology

Problem: Use a statistical spreadsheet or calculator to determine the population mean, median, and mode of the student test score data in Table 1 on page 108.

Approach: We will use Excel to obtain the mean, median, and mode. The steps for calculating measures of central tendency using the TI-83/84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step by Step on page 123.

Result: Figure 4 shows the output obtained from Excel. The #N/A in the output indicates that the data set has no mode.

Figure 4



```
                      Student Scores
Mean                              79
Standard Error           3.272783389
Median                          79.5
Mode                            #N/A
```

4 Use the Mean and the Median to Help Identify the Shape of a Distribution

Table 4

Mean	79
Median	79.5

Often, the mean and the median provide different values. Table 4 shows the mean and median scores on the exam for the data in Table 1 on page 108.

Notice that the median and the mean are close in value. Refer back to Table 1. Suppose Jennifer did not study for the exam and scored 28. The median would not change, but the mean would decrease from 79 to 75.6. We say that the median is **resistant** to extreme values (very large or small), but the mean is not resistant. Therefore, when data sets have unusually large or small values relative to the entire set of data or when the distribution of the data is skewed, the median is the preferred measure of central tendency over the mean because it is more representative of the typical observation.

In fact, the mean and median can be useful in determining the shape of a distribution. It can be shown that, if a distribution is perfectly symmetric and has one mode, then the median will equal the mean (and the mode). So symmetric distributions will have a median and a mean that are close in value. If the mean is substantially larger than the median, the distribution will be skewed right. Do you know why? In distributions that are skewed right, a few data values are substantially larger than the others. These larger data values cause the mean to be inflated while having little, if any, effect on the median. Similarly, distributions that are skewed left will have a mean that is substantially smaller than the median. We summarize these ideas in Table 5 and Figure 5.

CAUTION

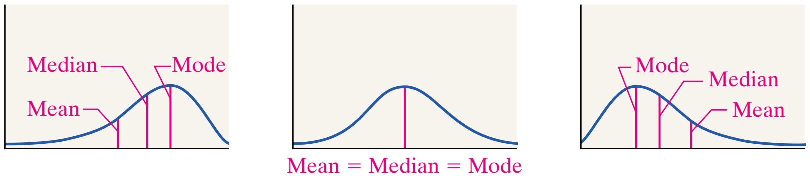
Because the mean is not resistant, it should not be reported as a measure of central tendency when the distribution of data is highly skewed.

Table 5

Relation between the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median

Figure 5
Mean/median versus skewness



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

EXAMPLE 8 Describing the Shape of a Distribution

Problem: In 2004, the New York Yankees had a record $184 million payroll. The data in Table 6 represent the salaries of the players on the opening-day roster in 2004 in thousands of dollars.

Table 6

Player	Salary	Player	Salary
Brown, Kevin	15,714	Lofton, Kenny	3,100
Cairo, Miguel	900	Matsui, Hideki	7,000
Clark, Tony	750	Mussina, Mike	16,000
Contreras, Jose	8,500	Osborne, Donovan	450
Crosby, Bubba	301	Posada, Jorge	9,000
De Paula, Jorge	303	Quantrill, Paul	3,000
Flaherty, John	775	Rivera, Mariano	10,890
Giambi, Jason	12,429	Rodriguez, Alex	22,000
Gordon, Tom	3,500	Sheffield, Gary	13,000
Heredia, Felix	1,800	Sierra, Ruben	1,000
Hernandez, Orlando	500	Vazquez, Javier	9,000
Jeter, Derek	18,600	White, Gabe	1,925
Karsay, Steve	6,000	Williams, Bernie	12,357
Lee, Travis	2,000	Wilson, Enrique	700
Lieber, Jon	2,700		

Source: usatoday.com

(a) Draw a frequency histogram of the Yankee salaries.

(b) Find the mean and median Yankee salary.

(c) Describe the shape of the distribution of Yankee salaries.

(d) Which measure of central tendency better describes the average salary of a player on the Yankees, the mean or the median?

Approach

(a) We will use MINITAB to draw a histogram of the salaries.

(b) We will use MINITAB to determine the mean and median salary.

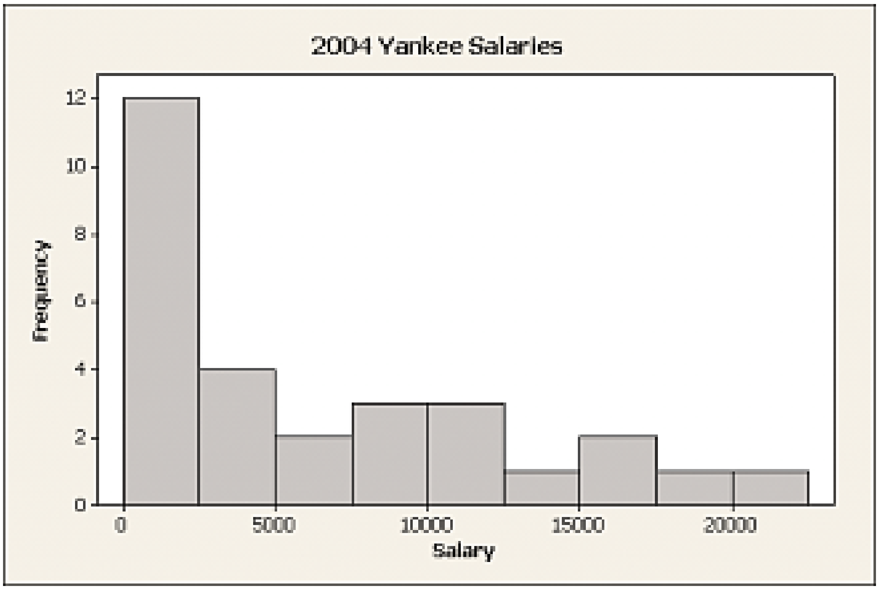
(c) We can identify the shape of the distribution by looking at the frequency histogram and comparing the mean to the median. Refer to Table 5 and Figure 5.

(d) If the data are skewed left or skewed right, the median is the better measure of central tendency. If the data are symmetric, the mean is the better measure of central tendency.

Solution

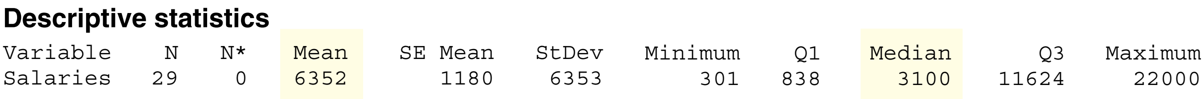
(a) Figure 6 shows a histogram of the data drawn using MINITAB.

Figure 6



(b) Using MINITAB, we find $\mu = 6352$ and $M = 3100$. See Figure 7.

Figure 7



```
Descriptive statistics
Variable   N   N*   Mean   SE Mean   StDev   Minimum    Q1   Median      Q3   Maximum
Salaries  29    0   6352      1180    6353       301   838     3100   11624     22000
```

(c) The shape of the histogram drawn in Figure 6 is skewed right. Notice that the mean is substantially larger than the median because the high salaries (especially Alex Rodriguez and Derek Jeter) push up the value of the mean.

(d) Because the shape of the distribution is skewed right, the median is the better measure of central tendency.

EXAMPLE 9 Describing the Shape of a Distribution

Problem: The data in Table 7 represent the birth weights (in pounds) of 50 randomly sampled babies.

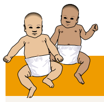


Table 7

5.8	7.4	9.2	7.0	8.5	7.6
7.9	7.8	7.9	7.7	9.0	7.1
8.7	7.2	6.1	7.2	7.1	7.2
7.9	5.9	7.0	7.8	7.2	7.5
7.3	6.4	7.4	8.2	9.1	7.3
9.4	6.8	7.0	8.1	8.0	7.5
7.3	6.9	6.9	6.4	7.8	8.7
7.1	7.0	7.0	7.4	8.2	7.2
7.6	6.7				

(a) Find the mean and the median.

(b) Describe the shape of the distribution.

(c) Which measure of central tendency better describes the average birth weight?

Approach

(a) Use a TI-84 Plus to compute the mean and the median.

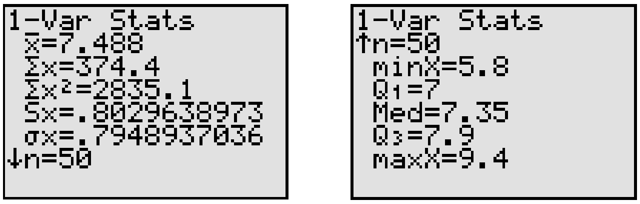
(b) The histogram, along with the mean and the median, is used to identify the shape of the distribution.

(c) If the data are roughly symmetric, the mean is the better measure of central tendency. If the data are skewed, the median is the better measure of central tendency.

Solution

(a) Using a TI-84 Plus, we find $\bar{x} = 7.49$ and $M = 7.35$. See Figure 8.

Figure 8



```
1-Var Stats
 x̄=7.488
 Σx=374.4
 Σx²=2835.1
 Sx=.8029638973
 σx=.7948937036
↓n=50
```

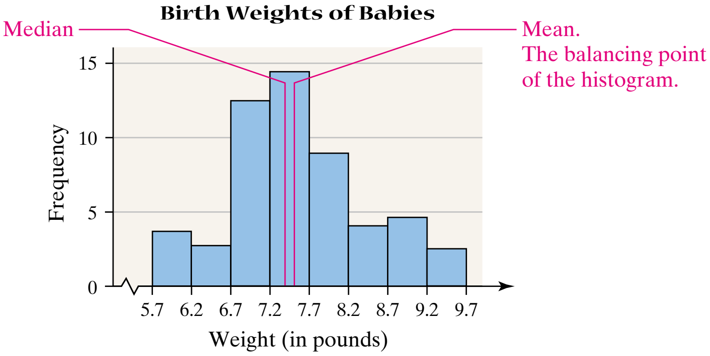
(a)

```
1-Var Stats
↑n=50
 minX=5.8
 Q₁=7
 Med=7.35
 Q₃=7.9
 maxX=9.4
```

(b)

(b) See Figure 9 for the frequency histogram with the mean and median labeled. The distribution is bell shaped. We have further evidence of the shape because the mean and median are close to each other.

Figure 9 Birth weights of 50 randomly selected babies



(c) Because the mean and median are close in value, we use the mean as the measure of central tendency.

Now Work Problem 29.

A question you may be asking yourself is, "Why would I ever compute the mean?" After all, the mean and median are close in value for symmetric data, and the median is the better measure of central tendency for skewed data. The reason we compute the mean is that much of the statistical inference that we perform is based on the mean. We will have more to say about this in Chapter 8.

We conclude this section with the following chart, which addresses the circumstances under which each measure of central tendency should be used.

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\Sigma x_i}{N}$ Sample mean: $\bar{x} = \frac{\Sigma x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and divide the data set in half	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or skewed right
Mode	Tally data to determine most frequent observation	Most frequent observation	When the most frequent observation is the desired measure of central tendency or the data are qualitative

3.1 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. What does it mean if a statistic is resistant? Why is the median resistant, but the mean is not? Is the mode a resistant measure of center?
2. Describe how the mean and the median can be used to determine the shape of a distribution.
3. In the 2000 census conducted by the U.S. Census Bureau, two average household incomes were reported: $41,349 and $55,263. One of these averages is the mean and the other is the median. Which is the mean? Support your answer.
4. The U.S. Department of Housing and Urban Development (HUD) uses the median to report the average price of a home in the United States. Why do you think HUD uses the median?
5. A histogram of a set of data indicates that the distribution of the data is skewed right. Which measure of central tendency will be larger, the mean or the median? Why?
6. If a data set contains 10,000 values arranged in increasing order, where is the median located?
7. Explain why the mode is used as the measure of central tendency for qualitative data.
8. A(n) _____ is a descriptive measure of a population, and a(n) _____ is a descriptive measure of a sample.
9. *True or False*: A data set will always have exactly one mode.
10. *True or False*: If the number of observations is odd, the median is $M = \frac{n+1}{2}$.

Skill Building

In Problems 11–14, find the population mean or sample mean as indicated.

11. Sample: 20, 13, 4, 8, 10
12. Sample: 83, 65, 91, 87, 84
13. Population: 3, 6, 10, 12, 14
14. Population: 1, 19, 25, 15, 12, 16, 28, 13, 6
15. For Super Bowl XXXIX, Fox television sold 59 ad slots for a total revenue of roughly $142 million. What was the mean price per ad slot?
16. The median for the given set of six ordered data values is 26.5. What is the missing value? 7 12 21 _____ 41 50
17. **Crash Test Results** The Insurance Institute for Highway Safety crashed the 2001 Honda Civic four times at 5 miles per hour. The costs of repair for each of the four crashes were

    $420, $462, $409, $236

    Compute the mean, median, and mode cost of repair.
18. **Cell Phone Use** The following data represent the monthly cell phone bill for my wife's phone for six randomly selected months.

    $35.34, $42.09, $39.43, $38.93 $43.39, $49.26

    Compute the mean, median, and mode phone bill.
19. **Concrete Mix** A certain type of concrete mix is designed to withstand 3000 pounds per square inch (psi) of pressure. The strength of concrete is measured by pouring the mix into casting cylinders 6 inches in diameter and 12 inches tall. The

NW

cylinder is allowed to "set up" for 28 days. The cylinders are then stacked on one another until the cylinders are crushed. The following data represent the strength of nine randomly selected casts (in psi).

3960, 4090, 3200, 3100, 2940, 3830, 4090, 4040, 3780

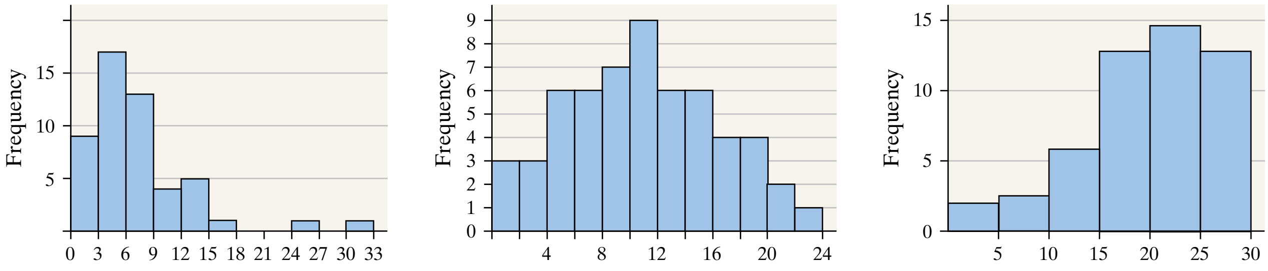
Compute the mean, median, and mode strength of the concrete (in psi).

20. **Flight Time** The following data represent the flight time (in minutes) of a random sample of seven flights from Las Vegas, Nevada, to Newark, New Jersey, on Continental Airlines.

282, 270, 260, 266, 257, 260, 267

Compute the mean, median, and mode flight time.

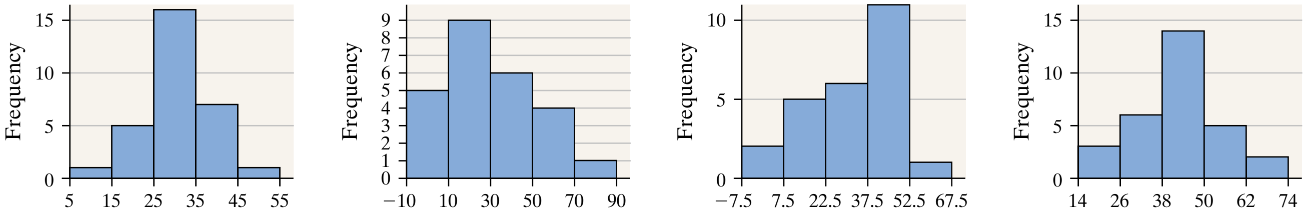
21. For each of the three histograms shown, determine whether the mean is greater than, less than, or approximately equal to the median. Justify your answer.



(a) **(b)** **(c)**

22. Match the histograms shown to the summary statistics shown to the right:

	Mean	Median
I	42	42
II	31	36
III	31	26
IV	31	32



(a) **(b)** **(c)** **(d)**

Applying the Concepts

23. **ATM Fees** The following data for a random sample of banks in Los Angeles and New York City represent the ATM fee for using another bank's ATM.

Los Angeles	2.00	1.50	1.50	1.00	1.50	2.00	0.00	2.00
New York City	1.50	1.00	1.00	1.25	1.25	1.50	1.00	0.00

Source: www.bankrate.com

Compute the mean, median, and mode ATM fee for each city. Does there appear to be a difference in the ATM fee charged in Los Angeles versus New York City? Why might this be the case?

24. **Reaction Time** In an experiment conducted online at the University of Mississippi, study participants are asked to react to a stimulus. In one experiment, the participant must press a key upon seeing a blue screen. The time (in seconds) to press the key is measured. The same person is then asked to press a key upon seeing a red screen, again with the time to react measured. The table shows the results for six study participants. Compute the mean, median, and mode reaction time for both blue and red. Does there appear to be a difference in the reaction time? What might account for any difference? How might this information be used?

Participant Number	Reaction Time to Blue	Reaction Time to Red
1	0.582	0.408
2	0.481	0.407
3	0.841	0.542
4	0.267	0.402
5	0.685	0.456
6	0.45	0.533

Source: PsychExperiments at the University of Mississippi (www.olemiss.edu/psychexps/)

25. Pulse Rates NW The following data represent the pulse rates (beats per minute) of nine students enrolled in a section of Sullivan's Introductory Statistics course. Treat the nine students as a population.

Student	Pulse
Perpectual Bempah	76
Megan Brooks	60
Jeff Honeycutt	60
Clarice Jefferson	81
Crystal Kurtenbach	72
Janette Lantka	80
Kevin McCarthy	80
Tammy Ohm	68
Kathy Wojdyla	73

(a) Compute the population mean pulse.
(b) Determine two simple random samples of size 3 and compute the sample mean pulse of each sample.
(c) Which samples result in a sample mean that overestimates the population mean? Which samples result in a sample mean that underestimates the population mean? Do any samples lead to a sample mean that equals the population mean?

26. Travel Time The following data represent the travel time (in minutes) to school for nine students enrolled in Sullivan's College Algebra course. Treat the nine students as a population.

Student	Travel Time	Student	Travel Time
Amanda	39	Scot	45
Amber	21	Erica	11
Tim	9	Tiffany	12
Mike	32	Glenn	39
Nicole	30		

(a) Compute the population mean for travel time.
(b) Determine three simple random samples of size 4 and compute the sample mean for travel time of each sample.
(c) Which samples result in a sample mean that overestimates the population mean? Which samples result in a sample mean that underestimates the population mean? Do any samples lead to a sample mean that equals the population mean?

27. Soccer Goals Mia Hamm, who retired after the 2004 Olympics, is considered by some to be the most prolific player in international soccer. The following data represent the number of goals scored over her 18-year career.

0	0	0	4	10	1	10	10	19
9	18	20	13	13	2	7	8	13

Source: www.soccerhall.com

(a) Compute the population mean of the number of goals she scored.
(b) Determine two simple random samples of size 3 and compute the sample mean of the number of goals she scored.
(c) Which samples result in a sample mean that overestimates the population mean? Which samples result in a sample mean that underestimates the population mean? Do any samples lead to a sample mean that equals the population mean?

28. Tour de Lance Lance Armstrong won the Tour de France seven consecutive times (1999–2005). The following table gives the winning times, distances, speeds, and margin of victory.

Year	Winning Time (h)	Distance (km)	Winning Speed (km/h)	Winning Margin (min)
1999	91.538	3687	40.28	7.617
2000	92.552	3662	39.56	6.033
2001	86.291	3453	40.02	6.733
2002	82.087	3278	39.93	7.283
2003	83.687	3427	40.94	1.017
2004	83.601	3391	40.56	6.317
2005	86.251	3593	41.65	44.667

Source: cyclingnews.com

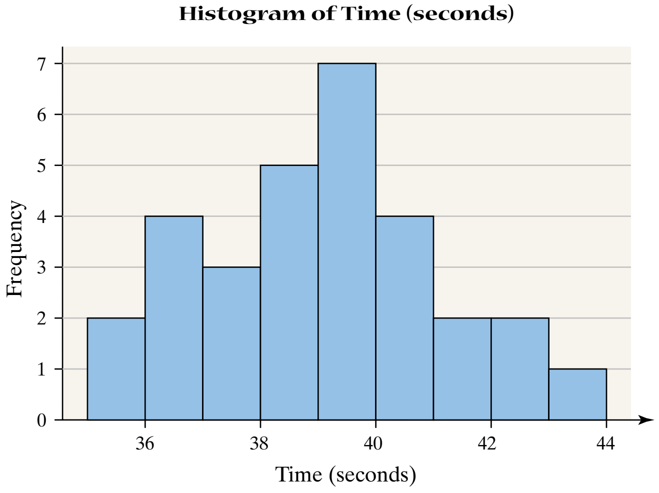
(a) Compute the mean and median of his winning times for the six races.
(b) Compute the mean and median of the distances for the six races.
(c) Compute the mean and median of his winning time margins.
(d) Compute the mean winning speed by finding the mean of the data values in the table. Next, compute the mean winning speed by finding the total of the six distances and dividing by the total of the six winning times. Finally, compute the mean winning speed by dividing the mean distance by the mean winning time. Do the three values agree or are there differences?

29. Connection Time NW The following data represent the connection time in seconds to an Internet service provider for 30 randomly selected connections.

39.76	36.13	36.61	38.80	39.04	39.09
37.24	35.62	40.07	38.76	39.23	38.38
38.24	36.34	35.89	42.86	36.03	37.03
38.64	41.86	41.22	37.19	40.50	39.81
39.84	39.45	40.91	43.12	40.54	42.02

Source: Nicole Spreitzer, student at Joliet Junior College

A histogram of the data is shown. The mean connection time is 39.007 seconds and the median connection time is 39.065 seconds. Use this information to identify the shape of the distribution. Which measure of central tendency better describes the "center" of the distribution?

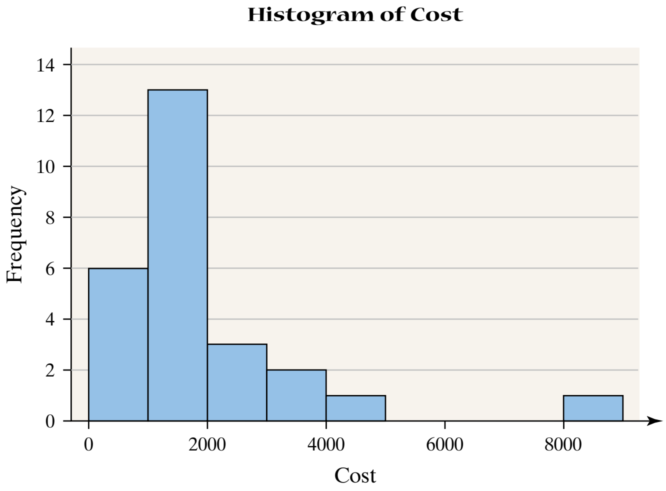

Histogram of Time (seconds)

30. **Journal Costs** The following data represent the annual subscription cost (in dollars) for a random sample of 26 biology journals.

1188	778	1970	661	1294	1175
2033	3911	198	8415	796	1840
1141	1050	3643	1407	1092	585
1049	1092	1589	4115	1150	2799
707	2330				

Source: Carol Wesolowski, student at Joliet Junior College

A histogram of the data is shown. The mean subscription cost is $1846 and the median subscription cost is $1182. Use this information to identify the shape of the distribution. Which measure of central tendency better describes the "center" of the distribution?


Histogram of Cost

31. **Serum HDL** Dr. Paul Oswiecmiski randomly selects 40 of his 20- to 29-year-old patients and obtains the following data regarding their serum HDL cholesterol.

70	56	48	48	53	52	66	48
36	49	28	35	58	62	45	60
38	73	45	51	56	51	46	39
56	32	44	60	51	44	63	50
46	69	53	70	33	54	55	52

(a) Compute the mean and the median serum HDL.
(b) Identify the shape of the distribution based on the histogram drawn in Problem 31 in Section 2.2 and the relationship between the mean and the median.

32. **Volume of Altria Group Stock** The volume of a stock is the number of shares traded on a given day. The following data represent the volume of Altria Group stock traded for a random sample of 35 trading days in 2004. The data are in millions, so 3.78 represents 3,780,000 shares traded.

3.78	8.74	4.35	5.02	8.40
6.06	5.75	5.34	6.92	6.23
5.32	3.25	6.57	7.57	6.07
3.04	5.64	5.00	7.16	4.88
10.32	3.38	7.25	6.52	4.43
3.38	5.53	4.74	9.70	3.56
10.96	4.50	7.97	3.01	5.58

Source: yahoo.finance.com

(a) Compute the mean and the median number of shares traded.
(b) Identify the shape of the distribution based on the histogram drawn in Problem 32 in Section 2.2 and the relationship between the mean and the median.

33. **M&Ms** NW The following data represent the weights (in grams) of a simple random sample of 50 M&M plain candies.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.94	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.93	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.76	0.90	
0.88	0.92	0.85	0.79	0.84	0.86	

Source: Michael Sullivan

Determine the shape of the distribution of weights of M&Ms by drawing a frequency histogram and computing the mean and median. Which measure of central tendency better describes the weight of a plain M&M?

34. Old Faithful We have all heard of the Old Faithful geyser in Yellowstone National Park. However, there is another, less famous, Old Faithful geyser in Calistoga, California. The following data represent the length of eruption (in seconds) for a random sample of eruptions of the California Old Faithful.

108	108	99	105	103	103	94
102	99	106	90	104	110	110
103	109	109	111	101	101	
110	102	105	110	106	104	
104	100	103	102	120	90	
113	116	95	105	103	101	
100	101	107	110	92	108	

Source: Ladonna Hansen, Park Curator

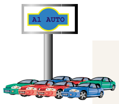
Determine the shape of the distribution of time between eruptions by drawing a frequency histogram and computing the mean and median. Which measure of central tendency better describes the time between eruptions?

35. Hours Working A random sample of 25 college students was asked, "How many hours per week typically do you work outside the home?" Their responses were as follows:

0	0	15	20	30
40	30	20	35	35
28	15	20	25	25
30	5	0	30	24
28	30	35	15	15

Determine the shape of the distribution of hours worked by drawing a frequency histogram and computing the mean and median. Which measure of central tendency better describes hours worked?

36. A Dealer's Profits The following data represent the profits (in dollars) of a new car dealer for a random sample of 40 sales.



781	1,038	453	1,446	3,082
501	451	1,826	1,348	3,001
1,342	1,889	580	0	2,909
2,883	480	1,664	1,064	2,978
149	1,291	507	261	540
543	87	798	673	2,862
1,692	1,783	2,186	398	526
730	2,324	2,823	1,676	4,148

Source: Ashley Hudson, student at Joliet Junior College

Determine the shape of the distribution of new car profits by drawing a frequency histogram and computing the mean and median. Which measure of central tendency better describes the profit?

37. Foreign-Born Population The following data represent the region of birth of foreign-born residents of the United States in 2003. Determine the mode region of birth.

Region	Number (thousands)
Caribbean	3,384
Central America	12,362
South America	2,111
Asia	8,375
Europe	4,590
Other Regions	2,680

Source: U.S. Census Bureau

38. Robbery The following data represent the number of offenses for various robberies in 2003. Determine the mode offense.

Type of Robbery	Number (thousands)
Street or highway	131
Commercial	61
Gas station	10
Convenience store	26
Residence	41
Bank	7

Source: U.S. Federal Bureau of Investigation

39. 2004 Presidential Election An exit poll was conducted in Los Alamos County, New Mexico, in which a random sample of 40 voters revealed whom they voted for in the presidential election. The results of the survey are shown below. Determine the mode candidate.

Kerry	Kerry	Bush	Bush	Bush
Bush	Kerry	Kerry	Bush	Bush
Kerry	Bush	Kerry	Bush	Kerry
Bush	Bush	Kerry	Kerry	Nader
Kerry	Bush	Bush	Kerry	Kerry
Badnarik	Kerry	Bush	Bush	Bush
Bush	Bush	Bush	Bush	Kerry
Kerry	Kerry	Kerry	Bush	Bush

40. Hospital Admissions The following data represent the diagnosis of a random sample of 20 patients admitted to a hospital. Determine the mode diagnosis.

Cancer	Motor vehicle accident	Congestive heart failure
Gunshot wound	fall	Gunshot wound
Gunshot wound	Motor vehicle accident	Gunshot wound
Assault	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Motor vehicle accident	Gunshot wound
Motor vehicle accident	Gunshot wound	Motor vehicle accident
Fall	Gunshot wound	

Source: Tamela Ohm, student at Joliet Junior College

41. Resistance and Sample Size Each of the following three data sets represents the IQ scores of a random sample of adults. IQ scores are known to have a mean and median of 100. For each data set, compute the mean and median. For each data set recalculate the mean and median, assuming that the individual whose IQ is 106 is accidentally recorded as 160. For each sample size, state what happens to the mean and the median. Comment on the role the number of observations plays in resistance.

Sample of Size 5

106	92	98	103	100

Sample of Size 12

106	92	98	103	100	102
98	124	83	70	108	121

Sample of Size 30

106	92	98	103	100	102
98	124	83	70	108	121
102	87	121	107	97	114
140	93	130	72	81	90
103	97	89	98	88	103

42. Super Bowl XXXIX Champion New England Patriots The following table gives roster information for the offense of the Super Bowl XXXIX Champion New England Patriots.

No.	Name	Position	Age (yr)	Weight (lb)	Years of Experience	College
12	Tom Brady	QB	27	225	5	Michigan
6	Rohan Davey	QB	26	245	3	LSU
13	Jim Miller	QB	33	225	10	Michigan State
27	Ribih Abdullah	RB	29	220	7	Lehigh
34	Cedric Cobbs	RB	24	225	Rookie	Arkansas
28	Corey Dillon	RB	30	225	8	Washington
33	Kevin Faulk	RB	28	202	6	LSU
35	Patrick Pass	RB	27	217	5	Georgia
83	Deion Branch	WR	25	193	3	Louisville
87	David Givens	WR	24	215	3	Notre Dame
81	Bethel Johnson	WR	25	200	2	Texas A&M
10	Kevin Kasper	WR	27	197	4	Iowa
86	David Patten	WR	30	190	8	Western Carolina
88	Christian Fauria	TE	33	250	10	Colorado
82	Daniel Graham	TE	26	257	3	Colorado
85	Jed Weaver	TE	28	258	6	Oregon
67	Daniel Koppen	C	25	296	2	Boston College
66	Lonie Paxton	C	26	260	5	Sacramento State
76	Brandon Gorin	OT	26	308	3	Purdue
72	Matt Light	OT	26	305	4	Purdue
63	Joe Andruzzi	OG	29	312	8	Southern Connecticut State
71	Russ Hochstein	OG	27	305	4	Nebraska
64	Gene Mruczkowski	OG	24	305	2	Purdue
61	Stephen Neal	OG	28	305	3	California State Bakersfield
74	Billy Yates	OG	24	305	2	Texas A&M

Source: ESPN.com

(a) Find the mean, median, and mode age.
(b) Find the mean, median, and mode weight.
(c) Find the mean, median, and mode years of experience. (**Note:** Rookie = 0 years.)
(d) Find the mode college attended.
(e) Obtain a simple random sample of six members of New England's offense. Compute sample mean age, weight, and years of experience. How do the sample means compare to the population means?
(f) Compute the mean, median, and mode weights of the five offensive guards (OG). Compute the mean, median, and mode weights of the five running backs (RB). Does there appear to be a difference in the weights? What might account for any differences?
(g) Does it make sense to compute the mean player number? Why?

43. Super Bowl XXXIX Champion New England Patriots Revisited Using the data presented in Problem 42, answer the following.

(a) The skilled positions in football are quarterback (QB), running back (RB), wide receiver (WR), and tight end (TE). Obtain a stratified sample by using skilled positions as one stratum and the remaining positions as a second stratum. Randomly select four skilled players and two "nonskilled" players. Compute the mean weight of the sample data. Compare to the result in part (e) in Problem 42.
(b) Cluster the players by position. Obtain a cluster sample by randomly selecting two clusters. Compute the mean weight of the sample data. Compare to the result in to part (e) in Problem 42. Can you think of any problems with obtaining a cluster sample?

44. You are negotiating a contract for the Players' Association of the NBA. Which measure of central tendency will you use to support your claim that the average player's salary needs to be increased? Why? As the chief negotiator for the owners, which measure would you use to refute the claim made by the Players' Association?

45. In January 2005, the mean amount of money lost per visitor to a local riverboat casino was $135. Do you think the median was more than, less than, or equal to this amount? Why?

46. Missing Exam Grade A professor has recorded exam grades for 20 students in his class, but one of the grades is no longer readable. If the mean score on the exam was 82 and the mean of the 19 readable scores is 84, what is the value of the unreadable score?

47. Suppose that the mean of a set of six data values is 34. What is the sum of the six data values?

48. For each of the following situations, determine which measure of central tendency is most appropriate and justify your reasoning.

(a) Average price of a home sold in Pittsburgh, Pennsylvania, in 2002
(b) Most popular major for students enrolled in a statistics course
(c) Average test score when the scores are distributed symmetrically
(d) Average test score when the scores are skewed right
(e) Average income of a player in the National Football League
(f) Most requested song at a radio station

49. Linear Transformations Benjamin owns a small Internet business. Besides himself, he employs nine other people. The salaries earned by the employees are given below in thousands of dollars (Benjamin's salary is the largest, of course):

30, 30, 45, 50, 50, 50, 55, 55, 60, 75

(a) Determine the mean, median, and mode for salary.
(b) Business has been good! As a result, Benjamin has a total of $25,000 in bonus pay to distribute to his employees. One option for distributing bonuses is to give each employee (including himself) $2500. Add the bonuses under this plan to the original salaries to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?
(c) As a second option, Benjamin can give each employee a bonus of 5% of his or her original salary. Add the bonuses under this second plan to the original salaries to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?
(d) As a third option, Benjamin decides not to give his employees a bonus at all. Instead, he keeps the $25,000 for himself. Use this plan to create a new data set. Recalculate the mean, median, and mode. How do they compare to the originals?

50. Linear Transformations Use the five test scores of 65, 70, 71, 75, and 95 to answer the following questions:

(a) Find the sample mean.
(b) Find the median.
(c) Which measure of central tendency best describes the typical test score?
(d) Suppose the professor decides to curve the exam by adding 4 points to each test score. Compute the sample mean based on the adjusted scores.
(e) Compare the unadjusted test score mean with the curved test score mean. What effect did adding 4 to each score have on the mean?

51. Trimmed Mean Another measure of central tendency is the trimmed mean. It is computed by determining the mean of a data set after deleting the smallest and largest observed values. Compute the trimmed mean for the data in Problem 33. Is the trimmed mean resistant? Explain.

52. Midrange The midrange is also a measure of central tendency. It is computed by adding the smallest and largest observed values of a data set and dividing the result by 2; that is,

$$\text{Midrange} = \frac{\text{largest data value} + \text{smallest data value}}{2}$$

Compute the midrange for the data in Problem 33. Is the midrange resistant? Explain.

Technology Step by Step **Determining the Mean and Median**

TI-83/84 Plus

***Step 1*:** Enter the raw data in L1 by pressing `STAT` and selecting `1:Edit`.

***Step 2*:** Press `STAT`, highlight the `CALC` menu, and select `1:1-Var Stats`.

***Step 3*:** With `1-Var Stats` appearing on the HOME screen, press `2nd 1` to insert `L1` on the `HOME` screen. Press `ENTER`.

MINITAB

***Step 1*:** Enter the data in C1.

***Step 2*:** Select the **Stat** menu, highlight **Basic Statistics**, and then highlight **Display Descriptive Statistics**.

***Step 3*:** In the **Variables** window, enter C1. Click OK.

Excel

***Step 1*:** Enter the data in column A.

***Step 2*:** Select the **Tools** menu and highlight **Data Analysis . . .**

***Step 3*:** In the Data Analysis window, highlight **Descriptive Statistics** and click OK.

***Step 4*:** With the cursor in the **Input Range** window, use the mouse to highlight the data in column A.

***Step 5*:** Select the **Summary statistics** option and click OK.

3.2 Measures of Dispersion

Objectives

1. Compute the range of a variable from raw data
2. Compute the variance of a variable from raw data
3. Compute the standard deviation of a variable from raw data
4. Use the Empirical Rule to describe data that are bell shaped
5. Use Chebyshev's Inequality to describe any set of data

In Section 3.1, we discussed measures of central tendency. The purpose of these measures is to describe the typical value of a variable. In addition to measuring the central tendency of a variable, we would also like to know the amount of dispersion in the variable. By *dispersion*, we mean the degree to which the data are "spread out." An example should help to explain why measures of central tendency are not sufficient in describing a distribution.

EXAMPLE 1 **Comparing Two Sets of Data**

***Problem*:** The data in Table 8 represent the IQ scores of a random sample of 100 students from two different universities. For each university, compute the mean IQ score and draw a histogram, using a lower class limit of 55 for the first class and a class width of 15. Comment on the results.

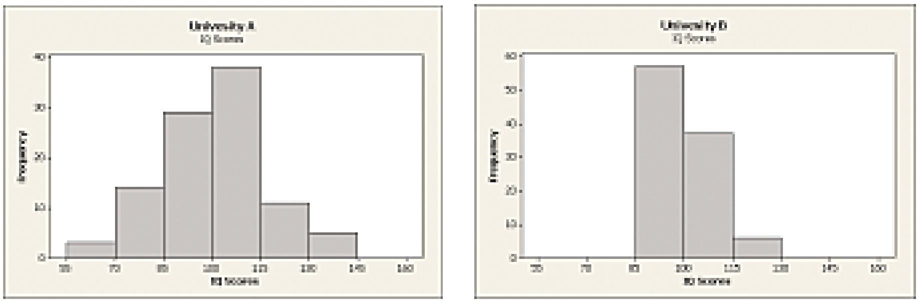
Table 8

University A										University B									
73	103	91	93	136	108	92	104	90	78	86	91	107	94	105	107	89	96	102	96
108	93	91	78	81	130	82	86	111	93	92	109	103	106	98	95	97	95	109	109
102	111	125	107	80	90	122	101	82	115	93	91	92	91	117	108	89	95	103	109
103	110	84	115	85	83	131	90	103	106	110	88	97	119	90	99	96	104	98	95
71	69	97	130	91	62	85	94	110	85	87	105	111	87	103	92	103	107	106	97
102	109	105	97	104	94	92	83	94	114	107	108	89	96	107	107	96	95	117	97
107	94	112	113	115	106	97	106	85	99	98	89	104	99	99	87	91	105	109	108
102	109	76	94	103	112	107	101	91	107	116	107	90	98	98	92	119	96	118	98
107	110	106	103	93	110	125	101	91	119	97	106	114	87	107	96	93	99	89	94
118	85	127	141	129	60	115	80	111	79	104	88	99	97	106	107	112	97	94	107

Approach: We will use MINITAB to compute the mean and draw a histogram for each university.

Solution: We enter the data into MINITAB and determine that the mean IQ score of both universities is 100.0. Figure 10 shows the histograms.

Figure 10



We notice that both universities have the same mean IQ, but the histograms indicate the IQs from University A are more spread out, that is, more dispersed. While an IQ of 100.0 is typical for both universities, it appears to be a more reliable description of the typical student from University B than from University A. That is, a higher proportion of students have IQ scores within, say, 15 points of the mean of 100.0 from University B than from University A.

Our goal in this section is to discuss numerical measures of dispersion so that we can quantify the spread of data. In this section, we discuss three numerical measures for describing the dispersion or spread of data: the range, variance, and standard deviation. In Section 3.4, we will discuss another measure of dispersion, the *interquartile range* (IQR).

1 Compute the Range of a Variable from Raw Data

The simplest measure of dispersion is the range. To compute the range, the data must be quantitative.

Definition

The **range**, ***R***, of a variable is the difference between the largest data value and the smallest data value. That is,

$$\text{Range} = R = \text{largest data value} - \text{smallest data value}$$

EXAMPLE 2 Computing the Range of a Set of Data

Table 9

Student	Score
1. Michelle	82
2. Ryanne	77
3. Bilal	90
4. Pam	71
5. Jennifer	62
6. Dave	68
7. Joel	74
8. Sam	84
9. Justine	94
10. Juan	88

Problem: The data in Table 9 represent the scores on the first exam of 10 students enrolled in a section of Introductory Statistics. Compute the range.

Approach: The range is found by computing the difference between the largest and smallest data values.

Solution: The highest test score is 94 and the lowest test score is 62. The range, R, is

$$94 - 62 = 32$$

All the students in the class scored between 62 and 94 on the exam. The difference between the best score and the worst score is 32 points.

Now compute the range of the data in Problem 19.

In Other Words
The range is not resistant.

Notice that the range is affected by extreme values in the data set, so the range is not resistant. If Jennifer did not study and scored 28, the range becomes $R = 94 - 28 = 66$. In addition, the range is computed using only two values in the data set (the largest and smallest). The *variance* and the *standard deviation*, on the other hand, use all the data values in the computations.

2 Compute the Variance of a Variable from Raw Data

Just as there is a population mean and sample mean, we also have a population variance and a sample variance. Measures of dispersion are meant to describe how spread out data are. Another way to think about this is to describe how far, on average, each observation is from the mean. Variance is based on the **deviation about the mean**. For a population, the deviation about the mean for the ith observation is $x_i - \mu$. For a sample, the deviation about the mean for the ith observation is $x_i - \bar{x}$. The further an observation is from the mean, the larger the absolute deviation.

The sum of all deviations about the mean must equal zero. That is,

$$\sum(x_i - \mu) = 0 \qquad \text{and} \qquad \sum(x_i - \bar{x}) = 0$$

In other words, observations larger than the mean are offset by observations smaller than the mean. Because the sum of deviations about the mean is zero, we cannot use the average deviation about the mean as a measure of spread. However, squaring a nonzero number always results in a positive number, so we could find the average squared deviation.

Definition

The **population variance** of a variable is the sum of the squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the mean of the squared deviations about the population mean. The population variance is symbolically represented by σ^2 (lowercase Greek sigma squared).

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N} = \frac{\sum(x_i - \mu)^2}{N} \qquad \textbf{(1)}$$

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

Note: In using Formula (1), do not round until the last computation. Use as many decimal places as allowed by your calculator to avoid round-off errors. ◀

A formula that is equivalent to Formula (1), called the **computational formula**, for determining the population variance is

$$\sigma^2 = \frac{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{N}}{N}$$

where Σx_i^2 means to square each observation and then sum these squared values, and $(\Sigma x_i)^2$ means to add up all the observations and then square the sum.

We illustrate how to use both formulas for computing the variance in the next example.

EXAMPLE 3 Computing a Population Variance

Problem: Compute the population variance of the test scores presented in Table 9.

Approach Using Formula (1)

Step 1: Create a table with four columns. Enter the population data in the first column. In the second column, enter the population mean.

Step 2: Compute the deviation about the mean for each data value. That is, compute $x_i - \mu$ for each data value. Enter these values in column 3.

Step 3: Square the values in column 3, and enter the results in column 4.

Step 4: Sum the squared deviations in column 4, and divide this result by the size of the population, N.

Solution

Step 1: See Table 10. Column 1 lists the observations in the data set, and column 2 contains the population mean.

Table 10

Score, x_i	**Population Mean, μ**	**Deviation about the Mean, $x_i - \mu$**	**Squared Deviations about the Mean, $(x_i - \mu)^2$**
82	79	$82 - 79 = 3$	$3^2 = 9$
77	79	$77 - 79 = -2$	$(-2)^2 = 4$
90	79	11	121
71	79	-8	64
62	79	-17	289
68	79	-11	121
74	79	-5	25
84	79	5	25
94	79	15	225
88	79	9	81
		$\sum(x_i - \mu) = 0$	$\sum(x_i - \mu)^2 = 964$

Step 2: Compute the deviations about the mean for each observation, as shown in column 3. For example, the deviation

Approach Using the Computational Formula

Step 1: Create a table with two columns. Enter the population data in the first column. Square each value in the first column and enter the result in the second column.

Step 2: Sum the entries in the first column. This is, find Σx_i. Sum the entries in the second column. That is, find Σx_i^2.

Step 3: Substitute the values found in Step 2 into the computational formula and simplify.

Solution

Step 1: See Table 11. Column 1 lists the observations in the data set, and column 2 contains the values in column 1 squared.

Table 11

Score, x_i	**Score Squared, x_i^2**
82	$82^2 = 6724$
77	$77^2 = 5929$
90	8100
71	5041
62	3844
68	4624
74	5476
84	7056
94	8836
88	7744
$\sum x_i = 790$	$\sum x_i^2 = 63{,}374$

Step 2: The last row of columns 1 and 2 shows that $\Sigma x_i = 790$ and $\Sigma x_i^2 = 63{,}374$.

about the mean for Michelle is 82 − 79 = 3. It is a good idea to add up the entries in this column to make sure they sum to 0.

***Step 3*:** Column 4 shows the squared deviations about the mean.

***Step 4*:** We sum the entries in column 4 to obtain the numerator of Formula (1). We compute the population variance by dividing the sum of the entries in column 4 by the number of students, 10:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{964}{10} = 96.4$$

***Step 3*:** We substitute 790 for Σx_i, 63,374 for Σx_i^2, and 10 for N into the computational formula.

$$\sigma^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}{N} = \frac{63{,}374 - \frac{(790)^2}{10}}{10}$$

$$= \frac{964}{10}$$

$$= 96.4$$

The unit of measure of the variance in Example 3 is points squared. This unit of measure results from squaring the deviations about the mean. Because points squared does not have any obvious meaning, the interpretation of variance is limited.

The sample variance is computed using sample data.

Definition

The **sample variance**, s^2, is computed by determining the sum of the squared deviations about the sample mean and dividing this result by $n - 1$. The formula for the sample variance from a sample of size n is

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n - 1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1} \qquad \textbf{(2)}$$

where $x_1, x_2, \ldots, x_n$ are the n observations in the sample and $\overline{x}$ is the sample mean.

CAUTION

When using Formula (2), be sure to use $\overline{x}$ with as many decimal places as possible to avoid round-off error.

A computational formula that is equivalent to Formula (2) for computing the sample variance is

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n - 1}$$

where Σx_i^2 means to square each observation and then sum these squared values, whereas $(\Sigma x_i)^2$ means to add up all the observations and then square the sum.

CAUTION

When computing the sample variance, be sure to divide by $n - 1$, not n.

Notice that the sample variance is obtained by dividing by $n - 1$. If we divided by n, as we might expect, the sample variance would consistently underestimate the population variance. Whenever a statistic consistently overestimates or underestimates a parameter, it is called **biased**. To obtain an unbiased estimate of the population variance, we divide the sum of the squared deviations about the mean by $n - 1$.

To help understand the idea of a biased estimator, consider the following situation: Suppose you work for a carnival in which you must guess a person's age. After 20 people come to your booth, you notice that you have a tendency to underestimate people's age. (You guess too low.) What would you do about this? In all likelihood, you would adjust your guesses higher so that you don't underestimate anymore. In other words, before the adjustment, your guesses were biased. To remove the bias, you increase your guess. That is what dividing by $n - 1$ in the sample variance formula accomplishes. Dividing by n results in an underestimate, so we divide by a smaller number to increase our "guess."

Although a proof that establishes why we divide by $n - 1$ is beyond the scope of the text, we can provide an explanation that has intuitive appeal. We already know that the sum of the deviations about the mean, $\Sigma(x_i - \overline{x})$, must equal zero. Therefore, if the sample mean is known and the first $n - 1$

observations are known, then the nth observation must be the value that causes the sum of the deviations to equal zero. For example, suppose $\bar{x} = 4$ based on a sample of size 3. In addition, if $x_1 = 2$ and $x_2 = 3$, then we can determine x_3.

$$\frac{x_1 + x_2 + x_3}{3} = \bar{x}$$

$$\frac{2 + 3 + x_3}{3} = 4 \qquad x_1 = 2, x_2 = 3, \bar{x} = 4$$

$$5 + x_3 = 12$$

$$x_3 = 7$$

We call $n - 1$ **degrees of freedom** because the first $n - 1$ observations have freedom to be whatever value they wish, but the nth value has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

Again, you should notice that Greek letters are used for parameters, while Roman letters are used for statistics. Do not use rounded values of the sample mean in Formula (2).

In Other Words

We have $n - 1$ degrees of freedom in the computation of s^2 because an unknown parameter, μ, is estimated with $\bar{x}$. For each parameter estimated, we lose 1 degree of freedom.

EXAMPLE 4 Computing a Sample Variance

Problem: Compute the sample variance of the sample obtained in Example 1(b) on page 108 from Section 3.1.

Approach: We follow the same approach that we used to compute the population variance, but this time using the sample data. In looking back at Example 1(b) from Section 3.1, we see that Bilal (90), Ryanne (77), Pam (71), and Michelle (82) are in the sample.

Solution Using Formula (2)

Step 1: Create a table with four columns. Enter the sample data in the first column. In the second column, enter the sample mean. See Table 12.

Table 12

Score, x_i	Sample Mean, $\bar{x}$	Deviation about the Mean, $x_i - \bar{x}$	Squared Deviations about the Mean, $(x_i - \bar{x})^2$
90	80	$90 - 80 = 10$	$10^2 = 100$
77	80	-3	9
71	80	-9	81
82	80	2	4
		$\sum(x_i - \bar{x}) = 0$	$\sum(x_i - \bar{x})^2 = 194$

Step 2: Compute the deviations about the mean for each observation, as shown in column 3. For example, the deviation about the mean for Bilal is $90 - 80 = 10$. It is a good idea to add up the entries in this column to make sure they sum to 0.

Step 3: Column 4 shows the squared deviations about the mean.

Step 4: We sum the entries in column 4 to obtain the numerator of Formula (1). We compute the population variance by dividing the sum of the entries in column 4 by one fewer than the number of students, $4 - 1$:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{194}{4 - 1} = 64.7$$

Solution Using the Computational Formula

Step 1: See Table 13. Column 1 lists the observations in the data set, and column 2 contains the values in column 1 squared.

Table 13

Score, x_i	Score Squared, x_i^2
90	$90^2 = 8{,}100$
77	$77^2 = 5{,}929$
71	5,041
82	6,724
$\sum x_i = 320$	$\sum x_i^2 = 25{,}794$

Step 2: The last rows of columns 1 and 2 show that $\Sigma x_i = 320$ and $\Sigma x_i^2 = 25{,}794$.

Step 3: We substitute 320 for Σx_i, 25,794 for Σx_i^2, and 4 for n into the computational formula.

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n - 1} = \frac{25{,}794 - \frac{(320)^2}{4}}{4 - 1}$$

$$= \frac{194}{3}$$

$$= 64.7$$

Notice that the sample variance obtained for this sample is an underestimate of the population variance we found in Example 3. This discrepancy does not violate our definition of an unbiased estimator, however. A biased estimator is one that *consistently* under- or overestimates.

3 Compute the Standard Deviation of a Variable from Raw Data

The standard deviation and the mean are the most popular methods for numerically describing the distribution of a variable. This is because these two measures are used for most types of statistical inference.

Definitions

The **population standard deviation**, $\boldsymbol{\sigma}$, is obtained by taking the square root of the population variance. That is,

$$\sigma = \sqrt{\sigma^2}$$

The **sample standard deviation**, $\boldsymbol{s}$, is obtained by taking the square root of the sample variance. That is,

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

EXAMPLE 5 Obtaining the Standard Deviation for a Population and a Sample

Problem: Use the results obtained in Examples 3 and 4 to compute the population and sample standard deviation score on the statistics exam.

Approach: The population standard deviation is the square root of the population variance. The sample standard deviation is the square root of the sample variance.

Solution: The population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}} = \sqrt{\frac{964}{10}} = 9.8 \text{ points}$$

The sample standard deviation for the sample obtained in Example 1 from Section 3.1 is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{194}{4 - 1}} = 8.0 \text{ points}$$

CAUTION
Never use the rounded variance to compute the standard deviation.

To avoid round-off error, never use the rounded value of the variance to compute the standard deviation.

Now Work Problem 25.

In-Class Activity: The Sample Standard Deviation

Using the pulse data from the activity in Section 3.1, page 109, do the following:

(a) Obtain a simple random sample of $n = 4$ students and compute the sample standard deviation.

(b) Obtain a second simple random sample of $n = 4$ students and compute the sample standard deviation.

(c) Are the sample standard deviations the same? Why?

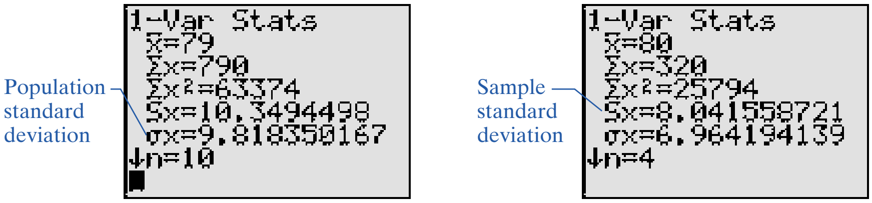
EXAMPLE 6 Determining the Variance and Standard Deviation Using Technology

Problem: Use a statistical spreadsheet or calculator to determine the population standard deviation of the data listed in Table 9. Also determine the sample standard deviation of the sample data from Example 4.

Approach: We will use a TI-84 Plus graphing calculator to obtain the population standard deviation and sample standard deviation score on the statistics exam. The steps for determining the standard deviation using the TI-83 or TI-84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step by Step on page 142.

Solution: Figure 11(a) shows the population standard deviation, and Figure 11(b) shows the sample standard deviation. Notice the TI graphing calculators provide both a population and sample standard deviation as output. This is because the calculator does not know whether the data entered are population data or sample data. It is up to the user of the calculator to choose the correct standard deviation. The results agree with those obtained in Example 5. To get the variance, we need to square the standard deviation. For example, the population variance is $9.818350167^2 = 96.4$ points^2.

Figure 11



```
1-Var Stats
 x̄=79
 Σx=790
 Σx²=63374
 Sx=10.3494498
 σx=9.818350167
↓n=10
```

(a)

```
1-Var Stats
 x̄=80
 Σx=320
 Σx²=25794
 Sx=8.041558721
 σx=6.964194139
↓n=4
```

(b)

Interpretations of the Standard Deviation

The standard deviation is used in conjunction with the mean to numerically describe distributions that are bell shaped and symmetric. The mean measures the center of the distribution, while the standard deviation measures the spread of the distribution. So how does the value of the standard deviation relate to the dispersion of the distribution? If we are comparing two populations, then **the larger the standard deviation, the more dispersion the distribution has**. This rule is true provided that the variable of interest from the two populations has the same unit of measure. The units of measure must be the same so that we are comparing apples with apples. For example, a standard deviation of $100 is not the same as 100 Japanese yen, because $1 is equivalent to about 109 yen. This means a standard deviation of $100 is substantially higher than a standard deviation of 100 yen.

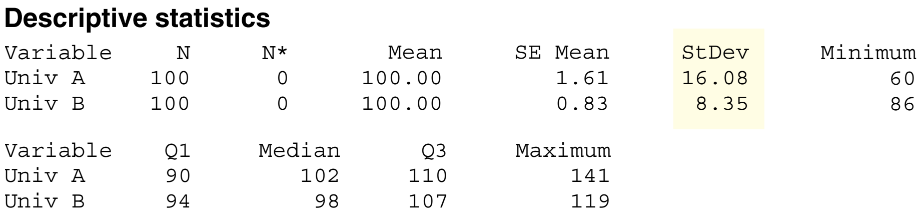
EXAMPLE 7 Comparing the Standard Deviation of Two Data Sets

Problem: Refer to the data in Example 1. Use the standard deviation to determine whether University A or University B has more dispersion in the IQ scores of its students.

Approach: We will use MINITAB to compute the standard deviation of IQ for each university. The university with the higher standard deviation will be the university with more dispersion in IQ scores. Recall that, on the basis of the histograms, it was apparent that University A had more dispersion. Therefore, we would expect University A to have a higher sample standard deviation.

Solution: We enter the data into MINITAB and compute the descriptive statistics. See Figure 12.

Figure 12
Descriptive Statistics



Descriptive statistics

```
Variable     N     N*      Mean    SE Mean   StDev    Minimum
Univ A     100      0    100.00       1.61   16.08         60
Univ B     100      0    100.00       0.83    8.35         86

Variable    Q1    Median     Q3    Maximum
Univ A      90       102    110        141
Univ B      94        98    107        119
```

The sample standard deviation is larger for University A (16.1) than for University B (8.4). Don't forget that we agreed to round the mean and standard deviation to one more decimal place than the original data. Therefore, University A has IQ scores that are more dispersed.

4 Use the Empirical Rule to Describe Data That Are Bell Shaped

If data have a distribution that is bell shaped, the following rule can be used to determine the percentage of data that will lie within k standard deviations of the mean.

The Empirical Rule

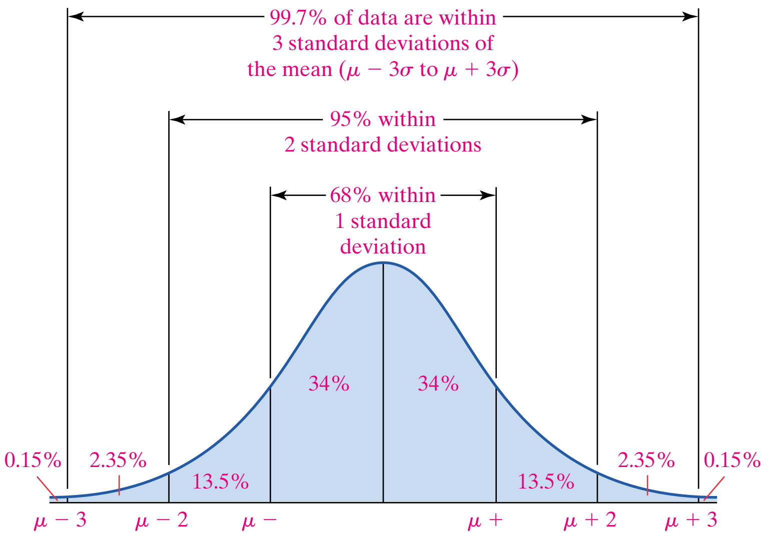
If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We can also use the Empirical Rule based on sample data with $\bar{x}$ used in place of μ and s used in place of σ. ◀

Figure 13 illustrates the Empirical Rule.

Figure 13



Let's revisit the data from University A in Table 8.

EXAMPLE 8 Using the Empirical Rule

Problem: Use the data from University A in Table 8.

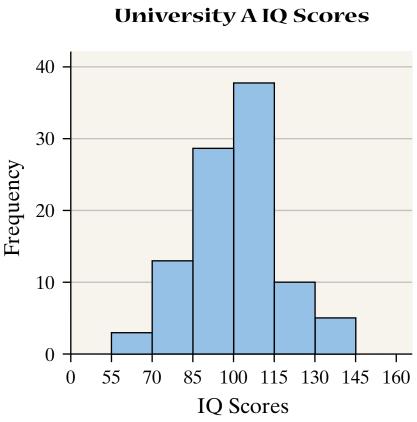
(a) Determine the percentage of students who have IQ scores within 3 standard deviations of the mean according to the Empirical Rule.

(b) Determine the percentage of students who have IQ scores between 67.8 and 132.2 according to the Empirical Rule.

(c) Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.

(d) According to the Empirical Rule, what percentage of students will have IQ scores above 132.2?

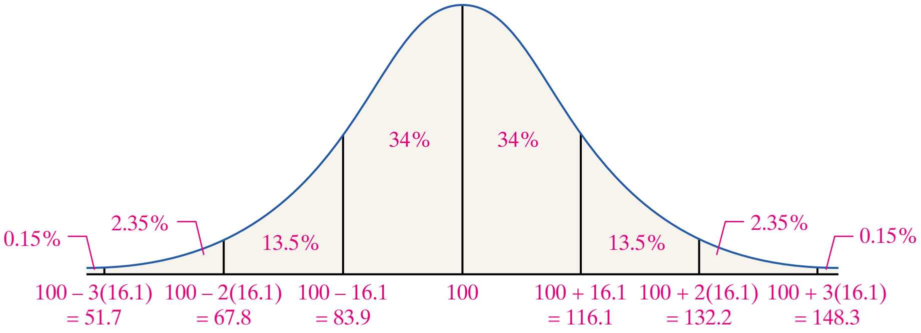
Figure 14



Approach: To use the Empirical Rule, a histogram of the data must be roughly bell shaped. Figure 14 shows the histogram of the data from University A.

Solution: The histogram of the data drawn in Figure 14 is roughly bell shaped. From Example 7 we know that the mean IQ score of the students enrolled in University A is 100 and the standard deviation is 16.1. To help organize our thoughts and make the analysis easier, we draw a bell-shaped curve like the one in Figure 13, with $\bar{x} = 100$ and $s = 16.1$. See Figure 15.

Figure 15



(a) According to the Empirical Rule, approximately 99.7% of the IQ scores will be within 3 standard deviations of the mean. That is, approximately 99.7% of the data will be greater than or equal to $100 - 3(16.1) = 51.7$ and less than or equal to $100 + 3(16.1) = 148.3$.

(b) Since 67.8 is exactly 2 standard deviations below the mean $[100 - 2(16.1) = 67.8]$ and 132.2 is exactly 2 standard deviations above the mean $[100 + 2(16.1) = 132.2]$, we use the Empirical Rule to determine that approximately 95% of all IQ scores lies between 67.8 and 132.2.

(c) Of the 100 IQ scores listed in Table 8, 96, or 96%, are between 67.8 and 132.2. This is very close to the approximation given by the Empirical Rule.

(d) Based on Figure 15, approximately $2.35\% + 0.15\% = 2.5\%$ of students at University A will have IQ scores above 132.2.

Now Work Problem 39.

5 Use Chebyshev's Inequality to Describe Any Set of Data

Chebyshev's Inequality was developed by the Russian mathematician Pafnuty Chebyshev (1821–1894). The inequality is used to determine a lower bound on the percentage of observations that lie within k standard deviations of the mean, where $k > 1$. What's amazing about this result is that these bounds are arrived at regardless of the basic shape of the distribution (skewed left, skewed right, or symmetric).

Chebyshev's Inequality

For any data set, regardless of the shape of the distribution, at least $\left(1 - \frac{1}{k^2}\right)100\%$ of the observations will lie within k standard deviations of the mean, where k is any number greater than 1. That is, at least $\left(1 - \frac{1}{k^2}\right)100\%$ of the data will lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

Note: We can also use Chebyshev's Inequality based on sample data.

For example, at least $\left(1 - \frac{1}{2^2}\right)100\% = 75\%$ of all observations will lie within $k = 2$ standard deviations of the mean and at least $\left(1 - \frac{1}{3^2}\right)100\% = 88.9\%$ of all observations will lie within $k = 3$ standard deviations of the mean.

Caution

The Empirical Rule holds only if the distribution is bell shaped. Chebyshev's Inequality holds regardless of the shape of the distribution.

Notice the result does not state that exactly 75% of all observations lie within 2 standard deviations of the mean, but instead states that 75% or more of the observations will lie within 2 standard deviations of the mean.

EXAMPLE 9 Using Chebyshev's Inequality

Problem: Using the data from University A in Table 8,

(a) Determine the minimum percentage of students who have IQ scores within 3 standard deviations of the mean according to Chebyshev's Inequality.

(b) Determine the minimum percentage of students who have IQ scores between 67.8 and 132.2, according to Chebyshev's Inequality.

(c) Determine the actual percentage of students who have IQ scores between 67.8 and 132.2.

Approach

(a) We use Chebyshev's Inequality with $k = 3$.

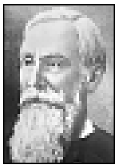
(b) We have to determine the number of standard deviations 67.8 and 132.2 are from the mean of 100.0. We then substitute this value of k into Chebyshev's Inequality.

(c) We refer to Table 8 and count the number of observations between 67.8 and 132.2. We divide this result by 100, the number of observations in the data set.

Solution

(a) We use Chebyshev's Inequality with $k = 3$ and determine that at least $\left(1 - \frac{1}{3^2}\right)100\% = 88.9\%$ of all students have IQ scores within 3 standard deviations of the mean. Since the mean of the data set is 100.0 and the standard deviation is 16.1, at least 88.9% of the students have IQ scores between $\bar{x} - ks = 100.0 - 3(16.1) = 51.7$ and $\bar{x} + ks = 100 + 3(16.1) = 148.3$.

(b) Since 67.8 is exactly 2 standard deviations below the mean $[100 - 2(16.1) = 67.8]$ and 132.2 is exactly 2 standard deviations above the mean $[100 + 2(16.1) = 132.2]$, we use Chebyshev's Inequality with $k = 2$



Historical Notes

Pafnuty Chebyshev was born on May 16, 1821, in Okatovo, Russia. In 1847, he began teaching mathematics at the University of St. Petersburg. Some of his more famous work was done on prime numbers. In particular, he discovered a way to determine the number of prime numbers less than or equal to a given number. Chebyshev also studied mechanics, including rotary motion. Chebyshev was elected a Fellow of the Royal Society in 1877. He died on November 26, 1894, in St. Petersburg.

to determine that at least $\left(1 - \frac{1}{2^2}\right)100\% = 75\%$ of all IQ scores lie between 67.8 and 132.2.

(c) Of the 100 IQ scores listed, 96 or 96% is between 67.8 and 132.2. Notice that Chebyshev's Inequality provides a rather conservative result.

Now Work Problem 43.

Because the Empirical Rule requires that the distribution be bell shaped, while Chebyshev's Inequality applies to all distributions, the Empirical Rule provides results that are more precise.

3.2 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. Would it be appropriate to say that a distribution with a standard deviation of 10 centimeters is more dispersed than a distribution with a standard deviation of 5 inches? Support your position.

2. What is meant by the phrase *degrees of freedom* as it pertains to the computation of the sample variance?

3. Is the standard deviation resistant?

4. The sum of the deviations about the mean always equals _____.

5. What does it mean when a statistic is biased?

6. The simplest measure of dispersion is the _____.

7. Discuss the relationship between variance and standard deviation.

8. The standard deviation is used in conjunction with the _____ to numerically describe distributions that are bell shaped. The _____ measures the center of the distribution, while the standard deviation measures the _____ of the distribution.

9. *True or False*: When comparing two populations, the larger the standard deviation, the more dispersion the distribution has, provided that the variable of interest from the two populations has the same unit of measure.

10. *True or False*: Chebyshev's Inequality applies to all distributions regardless of shape, but the Empirical Rule holds only for distributions that are bell shaped.

Skill Building

In Problems 11–16, find the population variance and standard deviation or the sample variance and standard deviation as indicated.

11. Sample: 20, 13, 4, 8, 10

12. Sample: 83, 65, 91, 87, 84

13. Population: 3, 6, 10, 12, 14

14. Population: 1, 19, 25, 15, 12, 16, 28, 13, 6

15. Sample: 6, 52, 13, 49, 35, 25, 31, 29, 31, 29

16. Population: 4, 10, 12, 12, 13, 21

17. Crash Test Results The Insurance Institute for Highway Safety crashed the 2001 Honda Civic four times at 5 miles per hour. The cost of repair for each of the four crashes is as follows:

$420, $462, $409, $236

Compute the range, sample variance, and sample standard deviation cost of repair.

18. Cell Phone Use The following data represent the monthly cell phone bill for my wife's phone for six randomly selected months:

$35.34, $42.09, $39.43, $38.93, $43.39, $49.26

Compute the range, sample variance, and sample standard deviation phone bill.

19. Concrete Mix NW A certain type of concrete mix is designed to withstand 3000 pounds per square inch (psi) of pressure. The strength of concrete is measured by pouring the mix into casting cylinders 6 inches in diameter and 12 inches tall. The cylinder is allowed to set up for 28 days. The cylinders are then stacked on one another until the cylinders are crushed. The following data represent the strength of nine randomly selected casts:

3960, 4090, 3200, 3100, 2940, 3830, 4090, 4040, 3780

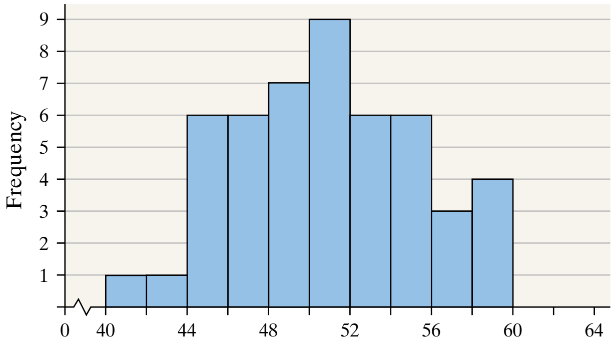
Compute the range, sample variance, and sample standard deviation for strength of the concrete (in psi).

20. Flight Time The following data represent the flight time (in minutes) of a random sample of seven flights from Las Vegas, Nevada, to Newark, New Jersey, on Continental Airlines.

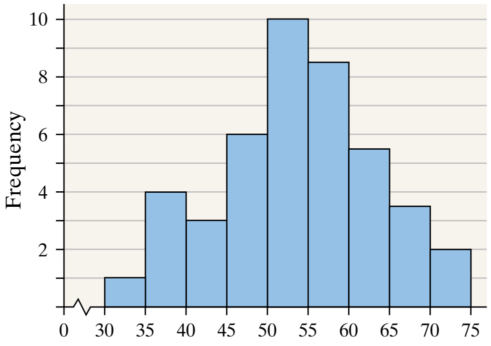
282, 270, 260, 266, 257, 260, 267

Compute the range, sample variance, and sample standard deviation of flight time.

21. Which histogram depicts a higher standard deviation? Justify your answer.



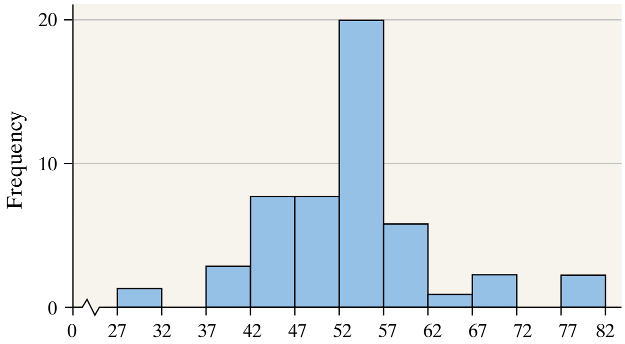
(a)



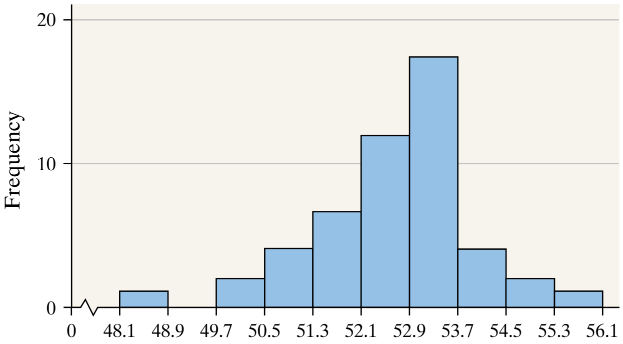
(b)

22. Match the histograms to the summary statistics given.

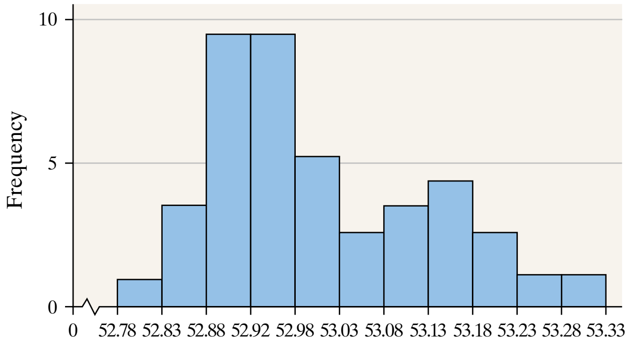
	Mean	Median	Standard Deviation
I	53	53	1.3
II	60	60	11
III	53	53	9
IV	53	53	0.12



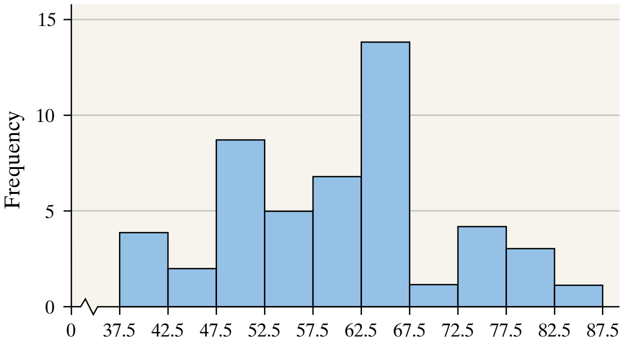
(a)



(b)



(c)



(d)

Applying the Concepts

23. ATM Fees The following data for a random sample of banks in Los Angeles and New York City represent the ATM fees for using another bank's ATM.

Los Angeles	2.00	1.50	1.50	1.00	1.50	2.00	0.00	2.00
New York City	1.50	1.00	1.00	1.25	1.25	1.50	1.00	0.00

Source: www.bankrate.com

Compute the range and sample standard deviation for ATM fees for each city. Which city has more dispersion based on the range? Which city has more dispersion based on the standard deviation?

24. Reaction Time In an experiment conducted online at the University of Mississippi, study participants are asked to react to a stimulus. In one experiment, the participant must press a key upon seeing a blue screen. The time (in seconds) to press the key is measured. The same person is then asked to press a key upon seeing a red screen, again with the time to react measured. The results for six study participants are listed in the table. Compute the range and sample standard deviation for reaction time for both blue and red. Does there appear to be a difference in the variability of reaction time? What might account for any difference?

Participant Number	Reaction Time to Blue	Reaction Time to Red
1	0.582	0.408
2	0.481	0.407
3	0.841	0.542
4	0.267	0.402
5	0.685	0.456
6	0.450	0.533

Source: PsychExperiments at the University of Mississippi (www.olemiss.edu/psychexps/)

25. Pulse Rates NW The following data represent the pulse rates (beats per minute) of nine students enrolled in a section of Sullivan's course in Introductory Statistics. Treat the nine students as a population.

Student	Pulse
Perpectual Bempah	76
Megan Brooks	60
Jeff Honeycutt	60
Clarice Jefferson	81
Crystal Kurtenbach	72
Janette Lantka	80
Kevin McCarthy	80
Tammy Ohm	68
Kathy Wojdyla	73

(a) Compute the population variance and population standard deviation.

(b) Determine two simple random samples of size 3, and compute the sample variance and sample standard deviation of each sample.

(c) Which samples underestimate the population standard deviation? Which overestimate the population standard deviation?

26. Travel Time The following data represent the travel time (in minutes) to school for nine students enrolled in Sullivan's College Algebra course. Treat the nine students as a population.

Student	Travel Time	Student	Travel Time
Amanda	39	Scot	45
Amber	21	Erica	11
Tim	9	Tiffany	12
Mike	32	Glenn	39
Nicole	30		

(a) Compute the population variance and population standard deviation.

(b) Determine three simple random samples of size 4, and compute the sample variance and sample standard deviation of each sample.

(c) Which samples underestimate the population standard deviation? Which overestimate the population standard deviation?

27. Soccer Goals Mia Hamm, considered by some to be the most prolific player in international soccer, retired after the 2004 Olympics. The following data represent the number of goals scored over her 18-year career.

0	0	0	4	10	1	10	10	19
9	18	20	13	13	2	7	8	13

Source: www.soccerhall.org

(a) Compute the population variance and population standard deviation for number of goals scored.

(b) Determine three simple random samples of size 3, and compute the sample variance and sample standard deviation of each sample.

(c) Which samples underestimate the population standard deviation? Which overestimate the population standard deviation?

28. Tour de Lance Lance Armstrong won the Tour de France seven consecutive times (1999–2005). The table gives the winning times, distances, speeds, and margin of victory.

Year	Winning Time (h)	Distance (km)	Winning Speed (km/h)	Winning Margin (min)
1999	91.538	3687	40.28	7.617
2000	92.552	3662	39.56	6.033
2001	86.291	3453	40.02	6.733
2002	82.087	3278	39.93	7.283
2003	83.687	3427	40.94	1.017
2004	83.601	3391	40.56	6.317
2005	86.251	3593	41.65	4.667

Source: www.cyclinynews.com

(a) Compute the range, population variance, and population standard deviation for winning times for the six races.

(b) Compute the range, population variance, and population standard deviation for distances for the six races.

(c) Compute the range, population variance, and population standard deviation for winning time margins.

(d) Compute the range, population variance, and population standard deviation for winning speeds.

29. A Fish Story Ethan and Drew went on a 10-day fishing trip. The number of smallmouth bass caught and released by the two boys each day was as follows:

Ethan:	9	24	8	9	5	8	9	10	8	10
Drew:	15	2	3	18	20	1	17	2	19	3

(a) Find the population mean and the range for the number of smallmouth bass caught per day by each fisherman. Do these values indicate any differences between the two fishermen's catches per day? Explain.

(b) Find the population standard deviation for the number of smallmouth bass caught per day by each fisherman. Do these values present a different story about the two fishermen's catches per day? Which fisherman has the more consistent record? Explain.

(c) Discuss limitations of the range as a measure of dispersion.

30. 2004 NFC Champion Philadelphia Eagles The following data represent the weights (in pounds) of the 33 offensive players and the 24 defensive players for the 2004 NFC Champion Philadelphia Eagles.

Offense			Defense		
195	218	215	281	272	265
240	222	212	265	264	298
210	250	200	303	293	306
243	205	180	294	240	241
210	195	226	254	245	262
180	245	258	200	196	194
255	244	312	210	177	210
300	305	310	211	206	202
340	330	327			
349	330	310			
320	330	325			

Source: ESPN.com

(a) Compute the population mean, the range, and the population standard deviation for the Philadelphia offense.

(b) Compute the population mean, the range, and the population standard deviation for the Philadelphia defense.

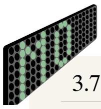
(c) Which player type has more dispersion? Explain how you know.

In Problems 31 and 32, compute the range, sample variance, and sample standard deviation.

31. Serum HDL Dr. Paul Oswiecmiski randomly selects 40 of his 20- to 29-year-old patients and obtains the following data regarding their serum HDL cholesterol:

70	56	48	48	53	52	66	48
36	49	28	35	58	62	45	60
38	73	45	51	56	51	46	39
56	32	44	60	51	44	63	50
46	69	53	70	33	54	55	52

32. Volume of Altria Group Stock The volume of a stock is the number of shares traded on a given day. The following data, given in millions so that 3.78 represents 3,780,000 shares traded, represent the volume of Altria Group stock traded for a random sample 35 trading days in 2004.



3.78	8.74	4.35	5.02	8.40
6.06	5.75	5.34	6.92	6.23
5.32	3.25	6.57	7.57	6.07
3.04	5.64	5.00	7.16	4.88
10.32	3.38	7.25	6.52	4.43
3.38	5.53	4.74	9.70	3.56
10.96	4.50	7.97	3.01	5.58

Source: Yahoo.finance.com

33. The Empirical Rule The following data represent the weights (in grams) of a random sample of 50 M&M plain candies.

0.87	0.88	0.82	0.90	0.90	0.84	0.84
0.91	0.94	0.86	0.86	0.86	0.88	0.87
0.89	0.91	0.86	0.87	0.93	0.88	
0.83	0.94	0.87	0.93	0.91	0.85	
0.91	0.91	0.86	0.89	0.87	0.93	
0.88	0.88	0.89	0.79	0.82	0.83	
0.90	0.88	0.84	0.93	0.76	0.90	
0.88	0.92	0.85	0.79	0.84	0.86	

Source: Michael Sullivan

(a) Determine the sample standard deviation weight. Express your answer rounded to two decimal places.
(b) On the basis of the histogram drawn in Section 3.1, Problem 33, comment on the appropriateness of using the Empirical Rule to make any general statements about the weights of M&Ms.
(c) Use the Empirical Rule to determine the percentage of M&Ms with weights between 0.79 and 0.95 gram. *Hint*: $\overline{x} = 0.87$.
(d) Determine the actual percentage of M&Ms that weigh between 0.79 and 0.95 gram, inclusive.
(e) Use the Empirical Rule to determine the percentage of M&Ms with weights more than 0.91 gram.
(f) Determine the actual percentage of M&Ms that weigh more than 0.91 gram.

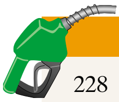
34. The Empirical Rule The following data represent the length of eruption for a random sample of eruptions at the Old Faithful geyser in Calistoga, California.

108	108	99	105	103	103	94
102	99	106	90	104	110	110
103	109	109	111	101	101	
110	102	105	110	106	104	
104	100	103	102	120	90	
113	116	95	105	103	101	
100	101	107	110	92	108	

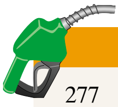
Source: Ladonna Hansen, Park Curator

(a) Determine the sample standard deviation length of eruption. Express your answer rounded to the nearest whole number.
(b) On the basis of the histogram drawn in Section 3.1, Problem 34, comment on the appropriateness of using the Empirical Rule to make any general statements about the length of eruptions.
(c) Use the Empirical Rule to determine the percentage of eruptions that last between 92 and 116 seconds. *Hint*: $\overline{x} = 104$.
(d) Determine the actual percentage of eruptions that last between 92 and 116 seconds, inclusive.
(e) Use the Empirical Rule to determine the percentage of eruptions that last less than 98 seconds.
(f) Determine the actual percentage of eruptions that last less than 98 seconds.

35. Which Car Would You Buy? Suppose that you are in the market to purchase a car. With gas prices on the rise, you have narrowed it down to two choices and will let gas mileage be the deciding factor. You decide to conduct a little experiment in which you put 10 gallons of gas in the car and drive it on a closed track until it runs out gas. You conduct this experiment 15 times on each car and record the number of miles driven.



Car 1				
228	223	178	220	220
233	233	271	219	223
217	214	189	236	248



Car 2				
277	164	326	215	259
217	321	263	160	257
239	230	183	217	230

Describe each data set. That is, determine the shape, center, and spread. Which car would you buy and why?

36. Which Investment Is Better? You have received a year-end bonus of $5000. You decide to invest the money in the stock market and have narrowed your investment options down to two mutual funds. The following data represent the historical quarterly rates of return of each mutual fund for the past 20 quarters (5 years).

Mutual Fund A				
1.3	−0.3	0.6	6.8	5.0
5.2	4.8	2.4	3.0	1.8
7.3	8.6	3.4	3.8	−1.3
6.4	1.9	−0.5	−2.3	3.1

Mutual Fund B				
−5.4	6.7	11.9	4.3	4.3
3.5	10.5	2.9	3.8	5.9
−6.7	1.4	8.9	0.3	−2.4
−4.7	−1.1	3.4	7.7	12.9

Describe each data set. That is, determine the shape, center, and spread. Which mutual fund would you invest in and why?

37. Rates of Return of Stocks Stocks may be categorized by industry. The following data represent the 5-year rates of return for a simple random sample of financial stocks and energy stocks ending March 4, 2004.

Financial Stocks

17.10	16.26	22.10	9.96	7.94	10.95	16.34	20.43
7.54	26.84	28.02	15.92	10.80	11.27	20.68	11.09
9.84	11.82	6.28	3.27	21.97	13.74	33.63	25.53
11.01	18.15	17.36	19.14	17.80	26.33	5.35	8.44

Energy Stocks

11.43	14.52	22.14	7.03	42.31	15.15	9.43	7.39
30.88	19.50	21.17	16.03	53.61	15.38	42.74	26.34
7.51	45.62	19.67	15.17	8.39	43.50	29.97	6.11
23.84	26.18	38.79	15.35	18.42	16.67	20.93	28.23

Source: Morningstar.com

(a) Compute the mean and the median rate of return for each industry. Which sector has the higher mean rate of return? Which sector has the higher median rate of return?

(b) Compute the standard deviation for each industry. In finance, the standard deviation rate of return is called **risk**. Which sector is riskier?

38. American League versus National League The following data represent the earned-run average of a random sample of pitchers in both the American League and the National League during the 2004 season. **Note**: Earned-run average (ERA) is the mean number of runs given up per nine innings pitched. A higher ERA is indicative of a worse pitcher.

American League

2.22	2.70	2.90	3.13	3.25
3.27	3.30	3.40	3.57	3.60
3.77	3.78	3.78	3.87	3.91
4.02	4.09	4.14	4.14	4.18
4.21	4.29	4.34	4.43	4.47
4.49	4.51	4.51	4.57	4.59
4.61	4.63	4.67	4.68	4.85
5.15	5.20	5.56	5.73	5.75

National League

2.34	2.39	2.43	2.84	2.95
3.09	3.11	3.20	3.20	3.24
3.28	3.54	3.59	3.62	3.68
3.76	3.78	3.81	3.82	3.84
3.87	3.96	3.99	4.00	4.01
4.08	4.11	4.11	4.13	4.13
4.16	4.19	4.23	4.30	4.43
4.45	4.52	4.52	4.59	4.64

(a) Compute the mean and the median earned-run average for each league. Which league has the higher mean earned-run average? Which league has the higher median earned-run average?

(b) Compute the standard deviation for each league. Which league has more dispersion?

39. The Empirical Rule One measure of intelligence is the NW Stanford–Binet Intelligence Quotient (IQ). IQ scores have a bell-shaped distribution with a mean of 100 and a standard deviation of 15.

(a) What percentage of people has an IQ score between 70 and 130?

(b) What percentage of people has an IQ score less than 70 or greater than 130?

(c) What percentage of people has an IQ score greater than 130?

40. The Empirical Rule SAT Math scores have a bell-shaped distribution with a mean of 518 and a standard deviation of 114. (*Source*: College Board, 2004)

(a) What percentage of SAT scores is between 404 and 632?

(b) What percentage of SAT scores is less than 404 or greater than 632?

(c) What percentage of SAT scores is greater than 746?

41. The Empirical Rule The weight, in grams, of the pair of kidneys in adult males between the ages of 40 and 49 has a bell-shaped distribution with a mean of 325 grams and a standard deviation of 30 grams.

(a) About 95% of kidneys will be between what weights?

(b) What percentage of kidneys weighs between 235 grams and 415 grams?

(c) What percentage of kidneys weighs less than 235 grams or more than 415 grams?

(d) What percentage of kidneys weighs between 295 grams and 385 grams?

42. The Empirical Rule The distribution of the length of bolts has a bell shape with a mean of 4 inches and a standard deviation of 0.007 inch.

(a) About 68% of bolts manufactured will be between what lengths?

(b) What percentage of bolts will be between 3.986 inches and 4.014 inches?

(c) If the company discards any bolts less than 3.986 inches or greater than 4.014 inches, what percentage of bolts manufactured will be discarded?

(d) What percentage of bolts manufactured will be between 4.007 inches and 4.021 inches?

43. Chebyshev's Inequality NW In December 2004, the average price of regular unleaded gasoline excluding taxes in the United States was $1.37 per gallon according to the Energy Information Administration. Assume that the standard deviation price per gallon is $0.05 per gallon to answer the following.

(a) What percentage of gasoline stations had prices within 3 standard deviations of the mean?

(b) What percentage of gasoline stations had prices within 2.5 standard deviations of the mean? What are the gasoline prices that are within 2.5 standard deviations of the mean?

(c) What is the minimum percentage of gasoline stations that had prices between $1.27 and $1.47?

44. Chebyshev's Inequality According to the U.S. Census Bureau, the mean of the commute time to work for a resident of Boston, Massachusetts, is 27.3 minutes. Assume that the standard deviation of the commute time is 8.1 minutes to answer the following:

(a) What percentage of commuters in Boston has a commute time within 2 standard deviations of the mean?

(b) What percentage of commuters in Boston has a commute time within 1.5 standard deviations of the mean? What are the commute times within 1.5 standard deviations of the mean?

(c) What is the minimum percentage of commuters who have commute times between 3 minutes and 51.6 minutes?

45. Comparing Standard Deviations The standard deviation of batting averages of all teams in the American League is 0.008. The standard deviation of all players in the American League is 0.02154. Why is there less variability in team batting averages?

46. Linear Transformations Benjamin owns a small Internet business. Besides himself, he employs nine other people. The salaries earned by the employees are given next in thousands of dollars (Benjamin's salary is the largest, of course):

30, 30, 45, 50, 50, 50, 55, 55, 60, 75

(a) Determine the range, population variance, and population standard deviation for the data.

(b) Business has been good! As a result, Benjamin has a total of $25,000 in bonus pay to distribute to his employees. One option for distributing bonuses is to give each employee (including himself) $2500. Add the bonuses under this plan to the original salaries to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?

(c) As a second option, Benjamin can give each employee a bonus of 5% of his or her original salary. Add the bonuses under this second plan to the original salaries to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?

(d) As a third option, Benjamin decides not to give his employees a bonus at all. Instead, he keeps the $25,000 for himself. Use this plan to create a new data set. Recalculate the range, population variance, and population standard deviation. How do they compare to the originals?

47. Resistance and Sample Size Each of the following three data sets represents the IQ scores of a random sample of adults. IQ scores are known to have a mean and median of 100. For each data set, determine the sample standard deviation. Then recompute the sample standard deviation assuming that the individual whose IQ is 106 is accidentally recorded as 160. For each sample size, state what happens to the standard deviation. Comment on the role that the number of observations plays in resistance.

Sample of Size 5				
106	92	98	103	100

Sample of Size 12					
106	92	98	103	100	102
98	124	83	70	108	121

Sample of Size 30					
106	92	98	103	100	102
98	124	83	70	108	121
102	87	121	107	97	114
140	93	130	72	81	90
103	97	89	98	88	103

48. Compute the sample standard deviation of the following test scores: 78, 78, 78, 78. What can be said about a data set in which all the values are identical?

Consumer Reports® BASEMENT WATERPROOFING COATINGS

A waterproofing coating can be an inexpensive and easy way to deal with leaking basements. But how effective are they? In a study, *Consumer Reports* tested nine waterproofers to rate their effectiveness in controlling water seepage though concrete foundations.

To compare the products' ability to control water seepage, we applied two coats of each product to slabs cut from concrete block. For statistical validity, this process was repeated at least six times. In each test run, four blocks (each coated with a different product) were simultaneously placed in a rectangular aluminum chamber. See the picture.

The chamber was sealed and filled with water and the blocks were subjected to progressively increasing hydrostatic pressures. Water that leaked out during each period was channeled to the bottom of the chamber opening, collected, and weighed.

The table contains a subset of the data collected for two of the products tested. Using these data,

(a) Calculate the mean, median, and mode weight of water collected for product A.

(b) Calculate the standard deviation of the weight of water collected for product A.

(c) Calculate the mean, median, and mode weight of water collected for product B.

(d) Calculate the standard deviation of the weight of water collected for product B.

(e) Construct a back-to-back stem-and-leaf diagram for these data.

Product	Replicate	Weight of Collected Water (in grams)
A	1	91.2
A	2	91.2
A	3	90.9
A	4	91.3
A	5	90.8
A	6	90.8
B	1	87.1
B	2	87.2
B	3	86.8
B	4	87.0
B	5	87.2
B	6	87.0

Does there appear to be a difference in these two products' ability to mitigate water seepage? Why?

Note to Readers: *In many cases, our test protocol and analytical methods are more complicated than described in these examples. The data and discussions have been modified to make the material more appropriate for the audience.*

Basement Waterproofer Test Chamber

Technology Step by Step **Determining the Range, Variance, and Standard Deviation**
The same steps followed to obtain the measures of central tendency from raw data can be used to obtain the measures of dispersion.

3.3 Measures of Central Tendency and Dispersion from Grouped Data

Preparing for This Section Before getting started, review the following:

- Organizing discrete data in tables (Section 2.2, pp. 71–72)
- Organizing continuous data in tables (Section 2.2, pp. 73–75)

Objectives

1 **Approximate the mean of a variable from grouped data**

2 **Compute the weighted mean**

3 **Approximate the variance and standard deviation of a variable from grouped data**

We have discussed how to compute descriptive statistics from raw data, but many times the data that we have access to have already been summarized in frequency distributions (grouped data). While we cannot obtain exact values of the mean or standard deviation without raw data, these measures can be approximated using the techniques discussed in this section.

1 Approximate the Mean of a Variable from Grouped Data

Since raw data cannot be retrieved from a frequency table, we assume that, within each class, the mean of the data values is equal to the *class midpoint*. The **class midpoint** is found by adding consecutive lower class limits and dividing the result by 2. We then multiply the class midpoint by the frequency. This product is expected to be close to the sum of the data that lie within the class. We repeat the process for each class and sum the results. This sum approximates the sum of all the data.

Definition

Approximate Mean of a Variable from a Frequency Distribution

Population Mean

$$\mu = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_N f_N}{f_1 + f_2 + \cdots + f_N} \qquad \textbf{(1a)}$$

Sample Mean

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \qquad \textbf{(1b)}$$

where x_i is the midpoint or value of the ith class
f_i is the frequency of the ith class
n is the number of classes

In Formula (1), $x_1 f_1$ approximates the sum of all the data values in the first class, $x_2 f_2$ approximates the sum of all the data values in the second class, and so on. Notice that the formulas for the population mean and sample mean are essentially identical, just as they were for computing the mean from raw data.

EXAMPLE 1 Approximating the Mean for Continuous Quantitative Data from the Frequency Distribution

Problem: The frequency distribution in Table 14 represents the three-year rate of return of a random sample of 40 small-capitalization growth mutual funds. Approximate the mean three-year rate of return.

Table 14

Class (3-year rate of return)	Frequency
0–1.99	2
2–3.99	5
4–5.99	6
6–7.99	8
8–9.99	9
10–11.99	6
12–13.99	3
14–15.99	1

Approach: We perform the following steps to approximate the mean.

Step 1: Determine the class midpoint of each class. The class midpoint is found by adding consecutive lower class limits and dividing the result by 2.

Step 2: Compute the sum of the frequencies, Σf_i.

Step 3: Multiply the class midpoint by the frequency to obtain $x_i f_i$ for each class.

Step 4: Compute $\Sigma x_i f_i$.

Step 5: Substitute into Formula (1b) to obtain the mean from grouped data.

Solution

Step 1: The lower class limit of the first class is 0. The lower class limit of the second class is 2. Therefore, the class midpoint of the first class is $\frac{0+2}{2} = 1$, so $x_1 = 1$. The remaining class midpoints are listed in column 2 of Table 15.

Step 2: We add the frequencies in column 3 to obtain $\Sigma f_i = 2 + 5 + \cdots + 1 = 40$.

Step 3: Compute the values of $x_i f_i$ by multiplying each class midpoint by the corresponding frequency and obtain the results shown in column 4 of Table 15.

Step 4: We add the values in column 4 of Table 15 to obtain $\Sigma x_i f_i = 304$.

Table 15

Class (3-year rate of return)	Class Midpoint, x_i	Frequency, f_i	$x_i f_i$
0–1.99	$\frac{0+2}{2} = 1$	2	$(1)(2) = 2$
2–3.99	3	5	$(3)(5) = 15$
4–5.99	5	6	30
6–7.99	7	8	56
8–9.99	9	9	81
10–11.99	11	6	66
12–13.99	13	3	39
14–15.99	15	1	15
		$\sum f_i = 40$	$\sum x_i f_i = 304$

Step 5: Substituting into Formula (1b), we obtain

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{304}{40} = 7.6$$

The approximate mean three-year rate of return is 7.6%.

The mean three-year rate of return from the raw data listed in Example 3 on page 74 from Section 2.2 is 7.5%. The approximate mean from grouped data is pretty close to the actual mean.

Note: To compute the mean from a frequency distribution where the data are discrete, treat each category of data as the class midpoint. For discrete data, the mean from grouped data will equal the mean from raw data. ◀

CAUTION

We computed the mean from grouped data in Example 1 even though the raw data are available. The reason for doing this was to illustrate how close the two values can be. In practice, use raw data whenever possible.

Now compute the mean of the frequency distribution in Problem 3.

2 Compute the Weighted Mean

Sometimes, certain data values have a higher importance or weight associated with them. In this case, we compute the *weighted mean*. For example, your grade-point average is a weighted mean, with the weights equal to the number of credit hours in each course. The value of the variable is equal to the grade converted to a point value.

Definition

The **weighted mean**, $\bar{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, summing these products, and dividing the result by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \qquad \textbf{(2)}$$

where w_i is the weight of the ith observation

x_i is the value of the ith observation

EXAMPLE 2 Computing the Weighted Mean

Problem: Marissa just completed her first semester in college. She earned an A in her 4-hour statistics course, a B in her 3-hour sociology course, an A in her 3-hour psychology course, a C in her 5-hour computer programming course, and an A in her 1-hour drama course. Determine Marissa's grade-point average.

Approach: We must assign point values to each grade. Let an A equal 4 points, a B equal 3 points, and a C equal 2 points. The number of credit hours for each course determines its weight. So a 5-hour course gets a weight of 5, a 4-hour course gets a weight of 4, and so on. We multiply the weight of each course by the points earned in the course, sum these products, and divide the sum by the number of credit hours.

Solution

$$\text{GPA} = \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{4(4) + 3(3) + 3(4) + 5(2) + 1(4)}{4 + 3 + 3 + 5 + 1} = \frac{51}{16} = 3.19$$

Marissa's grade-point average for her first semester is 3.19.

Now Work Problem 15.

3 Approximate the Variance and Standard Deviation of a Variable from Grouped Data

The procedure for approximating the variance and standard deviation from grouped data is similar to that of finding the mean from grouped data. Again, because we do not have access to the original data, the variance is approximate.

Definition

Approximate Variance of a Variable from a Frequency Distribution

Population Variance **Sample Variance**

$$\sigma^2 = \frac{\sum (x_i - \mu)^2 f_i}{\sum f_i} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\left(\sum f_i\right) - 1} \qquad \textbf{(3)}$$

where x_i is the midpoint or value of the ith class

f_i is the frequency of the ith class

An algebraically equivalent formula for the population variance is $\dfrac{\sum x_i^2 f_i - \dfrac{\left(\sum x_i f_i\right)^2}{\sum f_i}}{\sum f_i}$.

We approximate the standard deviation by taking the square root of the variance.

EXAMPLE 3 Approximating the Variance and Standard Deviation from a Frequency Distribution

Problem: The data in Table 14 on page 143 represent the three-year rate of return of a random sample of 40 small-capitalization growth mutual funds. Approximate the variance and standard deviation of the three-year rate of return.

Approach: We will use the sample variance Formula (3).

Step 1: Create a table with the class in the first column, the class midpoint in the second column, the frequency in the third column, and the unrounded mean in the fourth column.

Step 2: Compute the deviation about the mean, $x_i - \bar{x}$, for each class, where x_i is the class midpoint of the ith class and $\bar{x}$ is the sample mean. Enter the results in column 5.

Step 3: Square the deviation about the mean and multiply this result by the frequency to obtain $(x_i - \bar{x})^2 f_i$. Enter the results in column 6.

Step 4: Add the entries in columns 3 and 6 to obtain Σf_i and $\Sigma(x_i - \bar{x})^2 f_i$.

Step 5: Substitute the values obtained in Step 4 into Formula (3) to obtain an approximate value for the sample variance.

Solution

Step 1: We create Table 16. Column 1 contains the classes. Column 2 contains the class midpoint of each class. Column 3 contains the frequency of each class. Column 4 contains the unrounded sample mean obtained in Example 1.

Step 2: Column 5 of Table 16 contains the deviation about the mean, $x_i - \bar{x}$, for each class.

Step 3: Column 6 contains the values of the squared deviation about the mean multiplied by the frequency, $(x_i - \bar{x})^2 f_i$.

Step 4: We add the entries in columns 3 and 6 and obtain $\Sigma f_i = 40$ and $\Sigma(x_i - \bar{x})^2 f_i = 465.6$.

Table 16

Class (3-year rate of return)	Class Midpoint, x_i	Frequency, f_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
0–1.99	1	2	7.6	−6.6	87.12
2–3.99	3	5	7.6	−4.6	105.8
4–5.99	5	6	7.6	−2.6	40.56
6–7.99	7	8	7.6	−0.6	2.88
8–9.99	9	9	7.6	1.4	17.64
10–11.99	11	6	7.6	3.4	69.36
12–13.99	13	3	7.6	5.4	87.48
14–15.99	15	1	7.6	7.4	54.76
		$\sum f_i = 40$			$\sum(x_i - \bar{x})^2 f_i = 465.6$

Step 5: Substitute these values into Formula (3) to obtain an approximate value for the sample variance.

$$s^2 = \frac{\sum(x_i - \bar{x})^2 f_i}{\left(\sum f_i\right) - 1} = \frac{465.6}{39} \approx 11.94$$

Take the square root of the unrounded estimate of the sample variance to obtain an approximation of the sample standard deviation.

$$s = \sqrt{s^2} = \sqrt{\frac{465.6}{39}} \approx 3.46\%$$

We approximate the sample standard deviation three-year rate of return to be 3.46%.

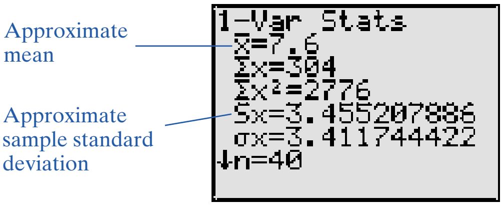
EXAMPLE 4 Approximating the Mean and Standard Deviation Using Technology

Problem: Approximate the mean and standard deviation of the three-year rate of return data in Table 14 using a TI-83/84 Plus graphing calculator.

Approach: The steps for approximating the mean and standard deviation of grouped data using the TI-83 or TI-84 Plus graphing calculator are given in the Technology Step by Step on page 149.

Result: Figure 16 shows the result from the TI-84 Plus.

Figure 16



```
1-Var Stats
 x̄=7.6
 Σx=304
 Σx²=2776
 Sx=3.455207886
 σx=3.411744422
↓n=40
```

From the output, we can see that the approximate mean is 7.6% and the approximate standard deviation is 3.46%. The results agree with our by-hand solutions.

From the raw data listed in Example 3, we find that the sample standard deviation is 3.46%. The approximate sample standard deviation from grouped data equals the sample standard deviation from the raw data!

Now compute the standard deviation from the frequency distribution in Problem 3.

3.3 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. Explain the role of the class midpoint in the formulas to approximate the mean and the standard deviation.

2. In Section 3.1, the mean is given by $\bar{x} = \frac{\Sigma x_i}{n}$. Explain how this is a special case of the weighted mean, $\bar{x}_w$.

Applying the Concepts

3. Cell Phones A sample of college students was asked how much they spent monthly on a cell phone plan (to the nearest NW dollar). Approximate the mean and standard deviation for the cost.

Monthly Cell Phone Plan Cost ($)	Number of Students
10–19	8
20–29	16
30–39	21
40–49	11
50–59	4

4. **Bowl Games** The following data represent the difference in scores between the winning and losing teams in the 2004–2005 college football bowl games. Approximate the mean and standard deviation for the point difference.

Point Difference	Number of Bowl Games
1–5	11
6–10	0
11–15	5
16–20	6
21–25	1
26–30	2
31–35	1
36–40	2

Source: espn.com

5. **100-Degree Days** The following data represent the annual number of days over 100°F for Dallas–Fort Worth from 1905 to 2004. Approximate the mean and standard deviation annual number of days over 100°F.

Number of 100° + Days	Number of Years
0–9	31
10–19	39
20–29	17
30–39	6
40–49	4
50–59	2
60–69	1

Source: National Weather Service

6. **Working Students** The following data represent the number of hours (on average) worked each week for a sample of community college students. Approximate the mean and standard deviation of the number of hours.

Hours Worked (per week)	Number of Students
0–9	24
10–19	14
20–29	39
30–39	18
40–49	5

7. **Health Insurance** The following data represent the number of people aged 25 to 64 years covered by health insurance (private or government) in 2003. Approximate the mean and standard deviation for age.

Age	Number (millions)
25–34	28.9
35–44	35.7
45–54	35.1
55–64	24.7

Source: U.S. Census Bureau

8. **Earthquakes** The following data represent the magnitude of earthquakes in the United States in 2004. Approximate the mean and standard deviation of the magnitude.

Magnitude	Number
0–0.9	539
1.0–1.9	1
2.0–2.9	1336
3.0–3.9	1363
4.0–4.9	289
5.0–5.9	21
6.0–6.9	2

Source: U.S. Geological Survey

9. **Meteorology** The following data represent the high-temperature distribution for the month of August in Chicago since 1872.

Temperature (°F)	Days
50–59	1
60–69	308
70–79	1519
80–89	1626
90–99	503
100–109	11

Source: National Oceanic and Atmospheric Administration

(a) Approximate the mean and standard deviation for temperature.
(b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.
(c) According to the Empirical Rule, 95% of days in the month of August will be between what two temperatures?

10. **Rainfall** The following data represent the annual rainfall distribution for St. Louis, Missouri, from 1870 to 2004.

Rainfall (inches)	Number of Years
20–24	4
25–29	15
30–34	27
35–39	40
40–44	28
45–49	15
50–54	4
55–59	2

Source: National Oceanic and Atmospheric Administration

(a) Approximate the mean and standard deviation for rainfall.
(b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.
(c) According to the Empirical Rule, 95% of annual rainfalls in St. Louis will be between what two amounts?

11. **Multiple Births** The following data represent the number of live multiple-delivery births (three or more babies) in 2002 for women 15 to 44 years old.

Age	Number of Multiple Births
15–19	93
20–24	511
25–29	1628
30–34	2832
35–39	1843
40–44	377

Source: *National Vital Statistics Reports*, Vol. 52, No. 10, December 17, 2003

(a) Approximate the mean and standard deviation for age.
(b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.
(c) According to the Empirical Rule, 95% of mothers of multiple births will be between what two ages?

12. **SAT Scores** The following data represent SAT Verbal scores for ISACS college-bound seniors in 2003.

SAT Verbal Score	Number
400–449	281
450–499	577
500–549	840
550–599	1120
600–649	1166
650–699	900
700–749	518
750–800	394

Source: www.isacs.org

(a) Approximate the mean and standard deviation of the score.
(b) Draw a frequency histogram of the data to verify that the distribution is bell shaped.
(c) According to the Empirical Rule, 95% of these ISACS college-bound seniors will have SAT Verbal scores between what two values?

13. **Serum HDL** Use the frequency distribution whose class width is 10 obtained in Problem 31 in Section 2.2 to approximate the mean and standard deviation for serum HDL. Compare these results to the values obtained in Problem 31 in Sections 3.1 and 3.2.

14. **Volume of Altria Group Stock** Use the frequency distribution whose class width is 2 obtained in Problem 32 in Section 2.2 to approximate the mean and standard deviation of the number of shares traded. Compare these results to the values obtained in Problem 32 in Sections 3.1 and 3.2.

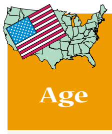
15. NW **Grade-Point Average** Marissa has just completed her second semester in college. She earned a B in her 5-hour calculus course, an A in her 3-hour social work course, an A in her 4-hour biology course, and a C in her 3-hour American literature course. Assuming that an A equals 4 points, a B equals 3 points, and a C equals 2 points, determine Marissa's grade-point average for the semester.

16. **Computing Class Average** In Marissa's calculus course, attendance counts for 5% of the grade, quizzes count for 10% of the grade, exams count for 60% of the grade, and the final exam counts for 25% of the grade. Marissa had a 100% average for attendance, 93% for quizzes, 86% for exams, and 85% on the final. Determine Marissa's course average.

17. **Mixed Chocolates** Michael and Kevin want to buy chocolates. They can't agree on whether they want chocolate-covered almonds, chocolate-covered peanuts, or chocolate-covered raisins. They agree to create a mix. They bought 4 pounds of chocolate-covered almonds at $3.50 per pound, 3 pounds of chocolate-covered peanuts for $2.75 per pound, and 2 pounds of chocolate-covered raisins for $2.25 per pound. Determine the cost per pound of the mix.

18. **Nut Mix** Michael and Kevin return to the candy store, but this time they want to purchase nuts. They can't decide among peanuts, cashews, or almonds. They again agree to create a mix. They bought 2.5 pounds of peanuts for $1.30 per pound, 4 pounds of cashews for $4.50 per pound, and 2 pounds of almonds for $3.75 per pound. Determine the price per pound of the mix.

19. **Population** The following data represent the male and female population by age of the United States for residents under 100 years old in July 2003.



Age	Male Resident Pop. (in thousands)	Female Resident Pop. (in thousands)
0–9	20,225	19,319
10–19	21,375	20,295
20–29	20,437	19,459
30–39	21,176	20,936
40–49	22,138	22,586
50–59	16,974	17,864
60–69	10,289	11,563
70–79	6,923	9,121
80–89	3,053	5,367
90–99	436	1,215

Source: U.S. Census Bureau

(a) Approximate the population mean and standard deviation of age for males.
(b) Approximate the population mean and standard deviation of age for females.
(c) Which gender has the higher mean age?
(d) Which gender has more dispersion in age?

20. Age of Mother The following data represent the age of the mother at childbirth for 1980 and 2002.

Age of Mother	Number of Births 1980 (thousands)	Number ofBirths 2002 (thousands)
10–14	1.1	0.7
15–19	53.0	43.0
20–24	115.1	103.6
25–29	112.9	113.6
30–34	61.9	91.5
35–39	19.8	41.4
40–44	3.9	8.3
45–49	0.2	0.5

Source: *National Vital Statistics Reports*, Vol. 52, No. 10

(a) Approximate the population mean and standard deviation of age for mothers in 1980.

(b) Approximate the population mean and standard deviation of age for mothers in 2002.

(c) Which year has the higher mean age?

(d) Which year has more dispersion in age?

Technology Step by Step

Determining the Mean and Standard Deviation from Grouped Data

TI-83/84 Plus

***Step 1*:** Enter the class midpoint in L1 and the frequency or relative frequency in L2 by pressing `STAT` and selecting `1:Edit`.

***Step 2*:** Press `STAT`, highlight the CALC menu and select `1:1-Var Stats`

***Step 3*:** With `1-Var Stats` appearing on the HOME screen, press `2nd 1` to insert `L1` on the `HOME` screen. Then press the comma and press `2nd 2` to insert `L2` on the `HOME` screen. So, the `HOME` screen should have the following:

```
1-Var Stats L1, L2
```

Press `ENTER` to obtain the mean and standard deviation.

3.4 Measures of Position

Objectives

1. Determine and interpret *z*-scores
2. Determine and interpret percentiles
3. Determine and interpret quartiles
4. Check a set of data for outliers

In Section 3.1, we were able to find measures of central tendency. Measures of central tendency are meant to describe the "typical" data value. Section 3.2 discussed measures of dispersion, which describe the amount of spread in a set of data. In this section, we discuss measures of position; that is, we wish to describe the relative position of a certain data value within the entire set of data.

1 Determine and Interpret *z*-Scores

At the end of the 2004 season, the Boston Red Sox led the American League with 949 runs scored, while the St. Louis Cardinals led the National League with 855 runs scored. A quick comparison might lead one to believe that the Red Sox are the better run-producing team. However, this comparison is unfair because the two teams play in different leagues. The Red Sox play in the American League, where the designated hitter bats for the pitcher, whereas the Cardinals play in the National League, where the pitcher must bat (pitchers are typically poor hitters). To compare the two teams' scoring of runs, we need to determine their relative standings in their respective leagues. This can be accomplished using a *z-score*.

Definition

The ***z*-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. It is obtained by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population *z*-score and a sample *z*-score; their formulas follow:

Population *z*-Score	Sample *z*-Score	
$z = \dfrac{x - \mu}{\sigma}$	$z = \dfrac{x - \overline{x}}{s}$	**(1)**

The *z*-score is unitless. It has mean 0 and standard deviation 1.

In Other Words

Z-scores provide a way to compare apples to oranges by converting variables with different centers and/or spreads to variables with the same center (0) and spread (1).

If a data value is larger than the mean, the *z*-score will be positive. If a data value is smaller than the mean, the *z*-score will be negative. If the data value equals the mean, the *z*-score will be zero. *Z*-scores measure the number of standard deviations an observation is above or below the mean. For example, a *z*-score of 1.24 is interpreted as "the data value is 1.24 standard deviations above the mean." A *z*-score of −2.31 is interpreted as "the data value is 2.31 standard deviations below the mean."

We are now prepared to determine whether the Red Sox or Cardinals had a better year in run production.

EXAMPLE 1 Comparing *z*-Scores

***Problem*:** Determine whether the Boston Red Sox or the St. Louis Cardinals had a relatively better run-producing season. The Red Sox scored 949 runs and play in the American League, where the mean number of runs scored was $\mu = 811.3$. The standard deviation was $\sigma = 73.7$. The Cardinals scored 855 runs and play in the National League, where the mean number of runs scored was $\mu = 751.1$. The standard deviation was $\sigma = 78.6$.

***Approach*:** To determine which team had the relatively better run-producing season, we compute each team's *z*-score. The team with the higher *z*-score had the better season. Because we know the values of the population parameters, we will compute the population *z*-score.

***Solution*:** First, we compute the *z*-score for the Red Sox. *Z*-scores are typically rounded to two decimal places.

$$z\text{-score} = \frac{x - \mu}{\sigma} = \frac{949 - 811.3}{73.7} = 1.87$$

Next, we compute the *z*-score for the Cardinals.

$$z\text{-score} = \frac{x - \mu}{\sigma} = \frac{855 - 751.1}{78.6} = 1.32$$

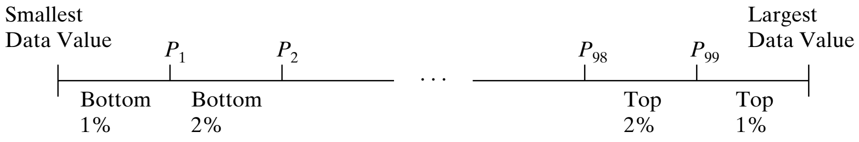
So, the Red Sox had run production 1.87 standard deviations above the mean, while the Cardinals had run production 1.32 standard deviations above the mean. Therefore, the Red Sox had a relatively better year at scoring runs.

Now Work Problem 7.

2 Determine and Interpret Percentiles

Recall that the median divides the lower 50% of a set of data from the upper 50%. In general, the ***k*th percentile**, denoted P_k, of a set of data divides the lower k% of a data set from the upper $(100 - k)$%. Percentiles divide a data set that is written in ascending order into 100 parts, so 99 possible percentiles can be computed. For example, P_1 divides the bottom 1% of the data from the top 99% while P_{99} divides the lower 99% of the data from the top 1%. Figure 17 displays the 99 possible percentiles.

Figure 17



If a data value lies at the 40th percentile, then approximately 40% of the data are less than this value and approximately 60% are higher than this value.

Percentiles are often used to give the relative standing of a data value. Many standardized exams, such as the SAT college entrance exam, use percentiles to provide students with an understanding of how they scored on the exam in relation to all other students who took the exam. For example, in 2004, an SAT verbal score of 580 was at the 73rd percentile. This means approximately 73% of the scores are below 580 and 27% are above 580. Pediatricians use percentiles to describe the progress of a newborn baby's weight gain relative to other newborn babies. A 3- to 5-month-old male child who weighs 14.3 pounds would be at the 15th percentile.

The following steps can be used to compute the kth percentile:

Determining the *k*th Percentile, P_k

***Step 1*:** Arrange the data in ascending order.

***Step 2*:** Compute an index i using the formula

$$i = \left(\frac{k}{100}\right)(n + 1) \qquad \textbf{(2)}$$

where k is the percentile of the data value and n is the number of individuals in the data set.

***Step 3*:**

(a) If i is an integer, the kth percentile, P_k, is the ith data value.

(b) If i is not an integer, find the mean of the observations on either side of i. This number represents the kth percentile, P_k.

CAUTION

Don't forget to write the data in ascending order before finding the percentile.

An example should clarify the procedure.

EXAMPLE 2 Determining the Percentile of a Data Value, Index an Integer

***Problem*:** The data in Table 17 represent the violent crime rate (violent crimes per 100,000 population) for the 50 states and the District of Columbia in 2003. Find the state that corresponds to the 75th percentile.

Table 17

State	Crime Rate	State	Crime Rate	State	Crime Rate
1. North Dakota	77.8	18. Oregon	295.5	35. Missouri	472.8
2. Maine	108.9	19. Connecticut	308.2	36. Oklahoma	505.7
3. Vermont	110.2	20. Mississippi	325.5	37. Michigan	511.2
4. New Hampshire	148.8	21. Ohio	333.2	38. Arizona	513.2
5. South Dakota	173.4	22. Colorado	345.1	39. Texas	552.5
6. Wisconsin	221.0	23. Washington	347.0	40. Illinois	556.8
7. Idaho	242.7	24. Indiana	352.8	41. California	579.3
8. Utah	248.6	25. Montana	365.2	42. Alaska	593.4
9. West Virginia	257.5	26. New Jersey	365.8	43. Nevada	614.2
10. Kentucky	261.7	27. Kansas	395.5	44. Louisiana	646.3
11. Wyoming	262.1	28. Pennsylvania	398.0	45. Delaware	658.0
12. Minnesota	262.6	29. Alabama	429.5	46. New Mexico	665.2
13. Hawaii	270.4	30. Georgia	453.9	47. Tennessee	687.8
14. Iowa	272.4	31. North Carolina	454.9	48. Maryland	703.9
15. Virginia	275.8	32. Arkansas	456.1	49. Florida	730.2
16. Rhode Island	285.6	33. New York	465.2	50. South Carolina	793.5
17. Nebraska	289.0	34. Massachusetts	469.4	51. District of Columbia	1608.1

Source: Federal Bureau of Investigation, Uniform Crime Reports, 2003

Approach: We will follow the steps given on page 151.

Solution

***Step 1*:** The data provided in Table 17 are already listed in ascending order.

***Step 2*:** To find the 75th percentile, P_{75}, we compute the index i with $k = 75$ and $n = 51$.

$$i = \left(\frac{75}{100}\right)(51 + 1) = 39$$

***Step 3*:** The 75th percentile is the 39th observation of the data set written in ascending order. The 39th observation, which corresponds to the state of Texas, is 552.5. Approximately 75% of the states have a violent crime rate less than 552.5 crimes per 100,000 population, and approximately 25% of the states have a violent crime rate above 552.5 crimes per 100,000 population.

EXAMPLE 3 Determining the Percentile of a Data Value, Index Not an Integer

Problem: Find the crime rate that corresponds to the 90th percentile for the data in Table 17.

Approach: We will follow the steps given on page 151.

Solution

***Step 1*:** The data provided in Table 17 are listed in ascending order.

***Step 2*:** To find the 90th percentile, P_{90}, we compute the index i with $k = 90$ and $n = 51$.

$$i = \left(\frac{90}{100}\right)(51 + 1) = 46.8$$

***Step 3*:** Because the index, $i = 46.8$, is not an integer, the 90th percentile is the mean of the 46th and 47th data value.

$$P_{90} = \frac{665.2 + 687.8}{2} = 676.5$$

Approximately 90% of the states have violent crime rates below 676.5 crimes per 100,000 population. Approximately 10% of the states have violent crime rates above 676.5 crimes per 100,000 population.

Now Work Problem 13(a).

Often we are interested in knowing the percentile to which a specific data value corresponds. The kth percentile of a data value, x, from a data set that contains n values is computed by using the following steps:

Finding the Percentile That Corresponds to a Data Value

***Step 1*:** Arrange the data in ascending order.

***Step 2*:** Use the following formula to determine the percentile of the score, x.

$$\text{Percentile of } x = \frac{\text{number of data values less than } x}{n} \times 100 \qquad \textbf{(3)}$$

Round this number to the nearest integer.

EXAMPLE 4 Finding the Percentile of a Specific Data Value

Problem: Find the percentile rank for the state of Kentucky using the data provided in Table 17.

Approach: We will follow the steps given above.

Solution

***Step 1*:** The data provided in Table 17 are in ascending order.

***Step 2*:** Nine states have a violent crime rate that is less than Kentucky's violent crime rate. So

$$\text{Percentile rank of Kentucky} = \frac{9}{51} \cdot 100 \approx 17.6$$

We round 17.6 to 18. Kentucky's violent crime rate is at the 18th percentile. Approximately 18% of the states have violent crime rates that are less than that of Kentucky, and approximately 82% of the states have violent crime rates that are larger than that of Kentucky.

Now Work Problem 13(d).

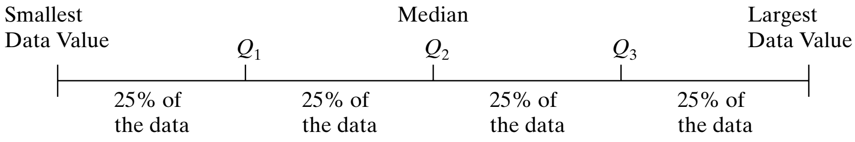
3 Determine and Interpret Quartiles

The most common percentiles are quartiles. **Quartiles** divide data sets into fourths, or four equal parts. The first quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile. The second quartile divides the bottom 50% of the data from the top 50%, so the second quartile is equivalent to the 50th percentile, which is equivalent to the median. Finally, the third quartile divides the bottom

75% of the data from the top 25%, so that the third quartile is equivalent to the 75th percentile. Figure 18 illustrates the concept of quartiles.

In Other Words
The first quartile, Q_1, is equivalent to the 25th percentile, P_{25}. The 2nd quartile, Q_2, is equivalent to the 50th percentile, P_{50}, which is equivalent to the median, M. Finally, the third quartile, Q_3, is equivalent to the 75th percentile, P_{75}.

Figure 18



EXAMPLE 5 Finding the Quartiles of a Data Set

Problem: Find the first, second, and third quartiles for the violent crime rates listed in Table 17.

Approach

Step 1: The first quartile, Q_1, is the 25th percentile, P_{25}. We let $k = 25$ in Formula (2) to obtain the index, i.

Step 2: The second quartile, Q_2, is the 50th percentile, P_{50}. We let $k = 50$ in Formula (2) to obtain the index, i.

Step 3: The third quartile, Q_3, is the 75th percentile, P_{75}. We let $k = 75$ in Formula (2) to obtain the index, i.

In Other Words
To find Q_2, determine the median of the data set. To find Q_1, determine the median of the "lower half" of the data set. To find Q_3, determine the median of the "upper half" of the data set.

Solution

Step 1: The index for the first quartile, Q_1, is

$$i = \left(\frac{25}{100}\right)(51 + 1) = 13$$

The 13th observation will be the first quartile. So $Q_1 = P_{25} = 270.4$.

Step 2: The index for the second quartile, Q_2, is

$$i = \left(\frac{50}{100}\right)(51 + 1) = 26$$

The 26th observation will be the second quartile. So $Q_2 = P_{50} = M = 365.8$

Step 3: The third quartile, Q_3, is the 75th percentile, which we found in Example 2. The third quartile is $Q_3 = P_{75} = 552.5$.

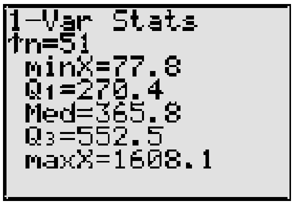
EXAMPLE 6 Finding Quartiles Using Technology

Problem: Find the quartiles of the violent crimes data in Table 17.

Approach: We will use a TI-84 Plus graphing calculator to obtain the quartiles. The steps for obtaining quartiles using a TI-83/84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step by Step on page 159.

Using Technology
Statistical packages may use different formulas for obtaining the quartiles, so results may differ slightly if the index is not an integer.

Figure 19



```
1-Var Stats
↑n=51
 minX=77.8
 Q1=270.4
 Med=365.8
 Q3=552.5
 maxX=1608.1
```

Result: Figure 19 shows the results obtained from a TI-84 Plus graphing calculator. Notice that the calculator uses Med for the second quartile.

Now Work Problem 15(b).

4 Check a Set of Data for Outliers

CAUTION

Outliers distort both the mean and the standard deviation, because neither is resistant. Because these measures often form the basis for most statistical inference, any conclusions drawn from a set of data that contains outliers can be flawed.

Whenever performing any type of data analysis, we should always check for extreme observations in the data set. Extreme observations are referred to as **outliers**. Whenever outliers are encountered, their origin must be investigated. They can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling. For example, in the 2000 presidential election, a precinct in New Mexico accidentally recorded 610 absentee ballots for Al Gore as 110. Workers in the Gore camp discovered the data-entry error through an analysis of vote totals.

Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European automobiles. If our sample included a Ferrari 360 Spider (approximately $170,000), it probably would be an outlier, because this car costs much more than the typical European automobile. The value of this car would be considered *unusual* because it is not a typical value from the data set.

We can use the following steps to check for outliers using quartiles.

Checking for Outliers by Using Quartiles

***Step 1*:** Determine the first and third quartiles of the data.

***Step 2*:** Compute the interquartile range. The **interquartile range** or **IQR** is the difference between the third and first quartile. That is,

$$\text{IQR} = Q_3 - Q_1$$

***Step 3*:** Determine the fences. **Fences** serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

***Step 4*:** If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

EXAMPLE 7 Checking for Outliers

***Problem*:** Check the data that represent the violent crime rates of the 50 states and the District of Columbia for outliers.

***Approach*:** We follow the preceding steps. Any data value that is less than the lower fence or greater than the upper fence will be considered an outlier.

Solution

***Step 1*:** The quartiles were found in Examples 5 and 6. So $Q_1 = 270.4$ and $Q_3 = 552.5$.

***Step 2*:** The interquartile range, IQR, is

$$\begin{aligned}\text{IQR} &= Q_3 - Q_1\\ &= 552.5 - 270.4\\ &= 282.1\end{aligned}$$

***Step 3*:** The lower fence, LF, is

$$\begin{aligned}\text{LF} &= Q_1 - 1.5(\text{IQR})\\ &= 270.4 - 1.5(282.1)\\ &= -152.75\end{aligned}$$

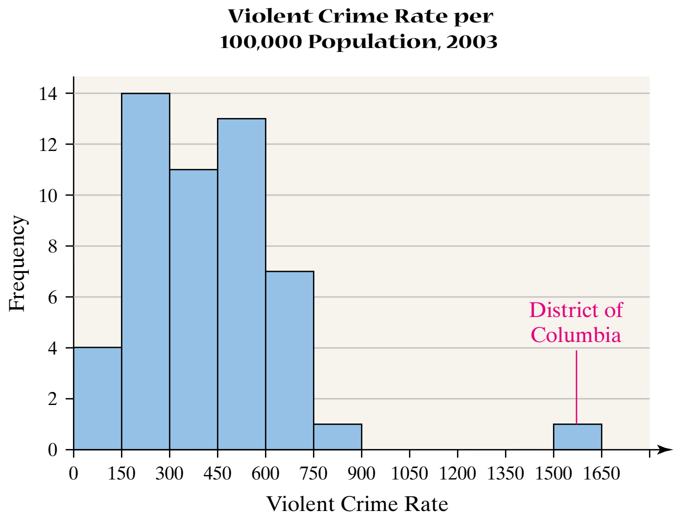
The upper fence, UF, is

$$\begin{aligned}\text{UF} &= Q_3 + 1.5(\text{IQR})\\ &= 552.5 + 1.5(282.1)\\ &= 975.65\end{aligned}$$

***Step 4*:** There are no outliers below the lower fence. However, we do have an outlier above the upper fence corresponding to the District of Columbia (1608.1 violent crimes per 100,000 population).

Figure 20 shows a histogram of the data. We can easily identify the outlier corresponding to the District of Columbia.

Figure 20



Now Work Problems 15(c) and (d).

3.4 ASSESS YOUR UNDERSTANDING

Concepts and Vocabulary

1. Write a paragraph that explains the meaning of percentiles.
2. Suppose you received the highest score on an exam. Your friend scored the second-highest score, yet you both were in the 99th percentile. How can this be?
3. Morningstar is a mutual fund rating agency. It ranks a fund's performance by using one to five stars. A one-star mutual fund is in the bottom 20% of its investment class; a five-star mutual fund is in the top 20% of its investment class. Interpret the meaning of a four-star mutual fund.
4. When outliers are discovered, should they always be removed from the data set before further analysis?
5. Mensa is an organization designed for people of high intelligence. One qualifies for Mensa if one's intelligence is measured at or above the 98th percentile. Explain what this means.
6. Explain the advantage of using z-scores to compare observations from two different data sets.

Applying the Concepts

7. **Birth Weights** In 2003, babies born after a gestation period of 32 to 35 weeks had a mean weight of 2600 grams and a standard deviation of 670 grams. In the same year, babies born after a gestation period of 40 weeks had a mean weight of 3500 grams and a standard deviation of 475 grams. Suppose a 34-week gestation period baby weighs 2400 grams and a 40-week gestation period baby weighs 3300 grams. Which baby weighs less relative to the gestation period?

NW

8. **Birth Weights** In 2003, babies born after a gestation per iod of 32 to 35 weeks had a mean weight of 2600 grams and a standard deviation of 670 grams. In the same year, babies born after a gestation period of 40 weeks had a mean weight of 3500 grams and a standard deviation of 475 grams. Suppose a 34-week gestation period baby weighs 3000 grams and a 40-week gestation period baby weighs 3900 grams. Which baby weighs less relative to the gestation period?

9. **Men versus Women** The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 2.7 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 2.6 inches. Who is relatively taller, a 75-inch man or a 70-inch woman?

(*Source*: *Vital and Health Statistics*, Advance Data, Number 347, October 27, 2004)

10. **Men versus Women** The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 2.7 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 2.6 inches. Who is relatively taller, a 68-inch man or a 62-inch woman?

(*Source*: *Vital and Health Statistics*, Advance Data, Oct. 2004)

11. **ERA Champions** In 2004, Jake Peavy of the San Diego Padres had the lowest ERA (earned-run average, mean number of runs yielded per nine innings pitched) of any pitcher in the National League, with an ERA of 2.27. Also in 2004, Johann Santana of the Minnesota Twins had the lowest ERA of any pitcher in the American League with an ERA of 2.61. In the National League, the mean ERA in 2004 was 4.198 and the standard deviation was 0.772. In the American League, the mean ERA in 2004 was 4.338 and the standard deviation was 0.785. Which player had the better year relative to his peers, Peavy or Santana? Why?

12. **Batting Champions** The highest batting average ever recorded was by Ted Williams in 1941 when he hit 0.406. That year, the mean and standard deviation for batting average were 0.28062 and 0.03281. In 2004, Ichiro Suzuki was the American League batting champion, with a batting average of 0.372. In 2004, the mean and standard deviation for batting average were 0.26992 and 0.02154. Who had the better year relative to their peers, Williams or Suzuki? Why?

13. **Violent Crime Rates** Use the data in Table 17 regarding NW the violent crime rates in 2003 to answer the following:
   (a) Find and interpret the 40th percentile.
   (b) Find and interpret the 95th percentile.
   (c) Find and interpret the 10th percentile.
   (d) What is the percentile rank of the state of Florida?
   (e) What is the percentile rank of the state of California?

14. **Violent Crime Rates** Use the data in Table 17 regarding the violent crime rates in 2003 to answer the following:
   (a) Find and interpret the 30th percentile.
   (b) Find and interpret the 85th percentile.
   (c) Find and interpret the 5th percentile.
   (d) What is the percentile rank of the state of New Mexico?
   (e) What is the percentile rank of the state of Rhode Island?

15. **April Showers** The following data represent the number NW of inches of rain in Chicago, Illinois, during the month of April for 20 randomly selected years.

0.97	2.78	4.00	5.50
1.14	3.41	4.02	5.79
1.85	3.48	4.11	6.14
2.34	3.94	4.77	6.28
2.47	3.97	5.22	7.69

Source: NOAA, Climate Diagnostics Center

   (a) Compute the z-score corresponding to the rainfall in 1971 of 0.97 inch. Interpret this result.
   (b) Determine the quartiles.
   (c) Compute the interquartile range, IQR.
   (d) Determine the lower and upper fences. Are there any outliers, according to this criterion?

16. **Hemoglobin in Cats** The following data represent the hemoglobin (in g/dL) for 20 randomly selected cats.

5.7	8.9	9.6	10.6	11.7
7.7	9.4	9.9	10.7	12.9
7.8	9.5	10.0	11.0	13.0
8.7	9.6	10.3	11.2	13.4

Source: Joliet Junior College Veterinarian Technology Program

   (a) Compute the z-score corresponding to the hemoglobin of Blackie, 7.8 g/dL. Interpret this result.
   (b) Determine the quartiles.
   (c) Compute the interquartile range, IQR.
   (d) Determine the lower and upper fences. Are there any outliers, according to this criterion?

17. **Concentration of Dissolved Organic Carbon** The following data represent the concentration of organic carbon (mg/L) collected from organic soil.

22.74	29.8	27.1	16.51	6.51
8.81	5.29	20.46	14.9	33.67
30.91	14.86	15.91	15.35	9.72
19.8	14.86	8.09	17.9	18.3
5.2	11.9	14	7.4	17.5
10.3	11.4	5.3	15.72	20.46
16.87	15.42	22.49		

Source: Lisa Emili, Ph.D. candidate, University of Waterloo, Ontario

   (a) Compute the z-score corresponding to 20.46. Interpret this result.
   (b) Determine the quartiles.
   (c) Compute the interquartile range, IQR.
   (d) Determine the lower and upper fences. Are there any outliers, according to this criterion?

18. Concentration of Dissolved Organic Carbon The following data represent the concentration of organic carbon (mg/L) collected from mineral soil.

8.5	3.91	9.29	21	10.89
10.3	11.56	7	3.99	3.79
5.5	4.71	7.66	11.72	11.8
8.05	10.72	21.82	22.62	10.74
3.02	7.45	11.33	7.11	9.6
12.57	12.89	9.81	17.99	21.4
8.37	7.92	17.9	7.31	16.92
4.6	8.5	4.8	4.9	9.1
7.9	11.72	4.85	11.97	7.85
9.11	8.79			

Source: Lisa Emili, Ph.D. candidate, University of Waterloo, Ontario

(a) Compute the z-score corresponding to 17.99. Interpret this result.
(b) Determine the quartiles.
(c) Compute the interquartile range, IQR.
(d) Determine the lower and upper fences. Are there any outliers, according to this criterion?

19. Fraud Detection As part of its "Customers First" program, a cellular phone company monitors monthly phone usage. The goal of the program is to identify unusual use and alert the customer that their phone may have been used by an unscrupulous individual. The following data represent the monthly phone use in minutes of a customer enrolled in this program for the past 20 months.

346	345	489	358	471
442	466	505	466	372
442	461	515	549	437
480	490	429	470	516

The phone company decides to use the upper fence as the cutoff point for the number of minutes at which the customer should be contacted. What is the cutoff point?

20. Stolen Credit Card A credit card company decides to enact a fraud-detection service. The goal of the credit card company is to determine if there is any unusual activity on the credit card. The company maintains a database of daily charges on a customer's credit card. Any day when the card was inactive is excluded from the database. If a day's worth of charges appears unusual, the customer is contacted to make sure that the credit card has not been compromised. Use the following daily charges (rounded to the nearest dollar) to determine the amount the daily charges must exceed before the customer is contacted.

143	166	113	188	133
90	89	98	95	112
111	79	46	20	112
70	174	68	101	212

21. Student Survey of Income A survey of 50 randomly selected full-time Joliet Junior College students was conducted during the Fall 2005 semester. In the survey, the students were asked to disclose their weekly income from employment. If the student did not work, $0 was entered.

0	262	0	635	0
244	521	476	100	650
12,777	567	310	527	0
83	159	0	547	188
719	0	367	316	0
479	0	82	579	289
375	347	331	281	628
0	203	149	0	403
0	454	67	389	0
671	95	736	300	181

(a) Check the data set for outliers.
(b) Draw a histogram of the data and label the outliers on the histogram.
(c) Provide an explanation for the outliers.

22. Student Survey of Entertainment Spending A survey of 40 randomly selected full-time Joliet Junior College students was conducted in the Fall 2005 semester. In the survey, the students were asked to disclose their weekly spending on entertainment. The results of the survey are as follows:

21	54	64	33	65
22	39	67	54	22
115	7	80	59	20
36	10	12	101	1000
28	28	75	50	27
32	51	33	26	35
21	26	13	38	9
16	14	36	8	48

(a) Check the data set for outliers.
(b) Draw a histogram of the data and label the outliers on the histogram.
(c) Provide an explanation for the outliers.

23. Pulse Rate Use the results of Problem 25 in Sections 3.1 and 3.2 to compute the z-scores for all the students. Compute the mean and standard deviation of these z-scores.

24. Travel Time Use the results of Problem 26 in Sections 3.1 and 3.2 to compute the z-scores for all the students. Compute the mean and standard deviation of these z-scores.

Technology Step by Step **Determining Percentiles**

TI-83/84 Plus To compute the quartiles, follow the same steps given to compute the mean and median from raw data.

MINITAB MINITAB computes only the quartiles. Follow the same steps given to compute the mean and median from raw data.

Excel ***Step 1*:** Enter the raw data into column A.

***Step 2*:** With the data analysis Tool Pak enabled, select the **Tools** menu and highlight **Data Analysis. . . .**

***Step 3*:** Select **Rank and Percentile** from the Data Analysis window.

***Step 4*:** With the cursor in the **Input Range** cell, highlight the data. Press OK.

3.5 The Five-Number Summary and Boxplots

Objectives

1. **Compute the five-number summary**
2. **Draw and interpret boxplots**

Some aspects of statistical analysis attempt to verify a conjecture by means of observational studies or designed experiments. In other words, a theory is conjectured, and then data are collected to test the theory. For example, a dietitian might conjecture that exercise will lower an individual's cholesterol. The dietitian would carefully design an experiment that randomly divides study participants into two groups: the control group and the experimental group. She would impose a treatment (exercise or no exercise) on the two groups and then measure the effect on the response variable, cholesterol levels.

Another aspect of statistics looks at data to spot any interesting results that might be concluded from the data. In other words, rather than develop a theory and use data to support or disprove the theory, a researcher starts with data and looks for a theory. This area of statistics is referred to as **exploratory data analysis (EDA)**. The idea behind exploratory data analysis is to draw graphs of data and obtain measures of central tendency and spread to form some conjectures regarding the data.

Many of the methods of exploratory data analysis were developed by John Tukey (1915–2000). A complete presentation of the materials found in this chapter can be found in his text *Exploratory Data Analysis* (Addison-Wesley, 1977).

1 Compute the Five-Number Summary

Remember that the median is a measure of central tendency that divides the lower 50% from the upper 50% of the data. This particular measure of central tendency is resistant to extreme values and is the preferred measure of central tendency when data are skewed right or left.

The three measures of dispersion presented in Section 3.2 (range, variance, and standard deviation) are not resistant to extreme values. However, the interquartile range, $Q_3 – Q_1$, is resistant. It measures the spread of the data by determining the difference between the 25th and 75th percentiles. It is interpreted as the range of the middle 50% of the data. However, the median, Q_1, and Q_3 do not provide information about the tails of the distribution of the data. To get this information, we need to know the smallest and largest values in the data set.

The **five-number summary** of a set of data consists of the smallest data value, Q_1, the median, Q_3, and the largest data value. Symbolically, the five-number summary is presented as follows:

Five-Number Summary

MINIMUM Q_1 M Q_3 MAXIMUM

EXAMPLE 1 Obtaining the Five-Number Summary

Problem: The data shown in Table 18 show the finishing times (in minutes) of the men in the 60- to 64-year-old age group in a 5-kilometer race.

Table 18

19.95	23.25	23.32	25.55	25.83	26.28
28.58	28.72	30.18	30.35	30.95	32.13
33.23	33.53	36.68	37.05	37.43	41.42
42.47	49.17	64.63			

Source: Laura Gillogly, student at Joliet Junior College

Approach: The five-number summary requires that we determine the minimum data value, Q_1 (the 25th percentile), M (the median), Q_3 (the 75th percentile), and the maximum data value. We need to arrange the data in ascending order and then use the procedures introduced in Section 3.4 to obtain Q_1, M, and Q_3.

Solution: The data in ascending order are as follows:

19.95, 23.25, 23.32, 25.55, 25.83, 26.28, 28.58, 28.72, 30.18, 30.35, 30.95, 32.13, 33.23, 33.53, 36.68, 37.05, 37.43, 41.42, 42.47, 49.17, 64.63

The smallest number (the fastest time) in the data set is 19.95. The largest number in the data set is 64.63. The first quartile, Q_1, is 26.06. The median, M, is 30.95. The third quartile, Q_3, is 37.24. The five-number summary is

19.95 26.06 30.95 37.24 64.63

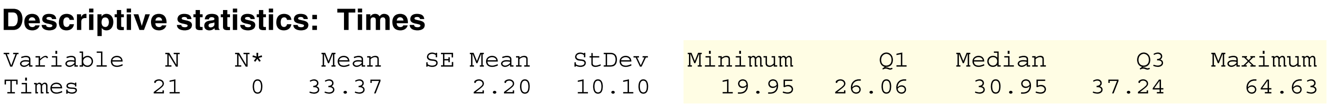
EXAMPLE 2 Obtaining the Five-Number Summary Using Technology

Problem: Using a statistical spreadsheet or graphing calculator, determine the five-number summary of the data presented in Table 18.

Approach: We will use MINITAB to obtain the five-number summary. The steps for obtaining the five-number summary using a TI-83 or TI-84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step by Step on page 168.

Result: Figure 21 shows the output supplied by MINITAB. The five-number summary is highlighted.

Figure 21



Descriptive statistics: Times

```
Variable    N    N*    Mean   SE Mean   StDev   Minimum       Q1    Median       Q3    Maximum
Times      21     0   33.37      2.20   10.10     19.95    26.06     30.95    37.24      64.63
```

2 Draw and Intepret Boxplots

The five-number summary can be used to create another graph, called the **boxplot**.

Drawing a Boxplot

***Step 1*:** Determine the lower and upper fences:

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Remember, IQR $= Q_3 - Q_1$.

***Step 2*:** Draw vertical lines at Q_1, M, and Q_3. Enclose these vertical lines in a box.

***Step 3*:** Label the lower and upper fences.

***Step 4*:** Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence.

***Step 5*:** Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*).

EXAMPLE 3 Constructing a Boxplot

***Problem*:** Use the results from Example 1 to a construct a boxplot of the finishing times of the men in the 60- to 64-year-old age group.

***Approach*:** Follow the steps presented above.

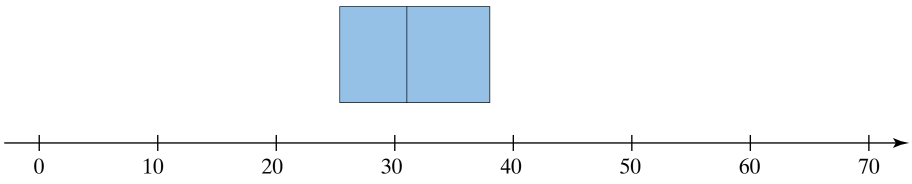
***Solution*:** From the results of Example 1, we know that $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Therefore, the interquartile range = IQR = $Q_3 - Q_1 = 37.24 - 26.06 = 11.18$. The difference between the 75th percentile and 25th percentile is a time of 11.18 minutes.

***Step 1*:** We compute the lower and upper fences:

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR}) = 26.06 - 1.5(11.18) = 9.29$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR}) = 37.24 + 1.5(11.18) = 54.01$$

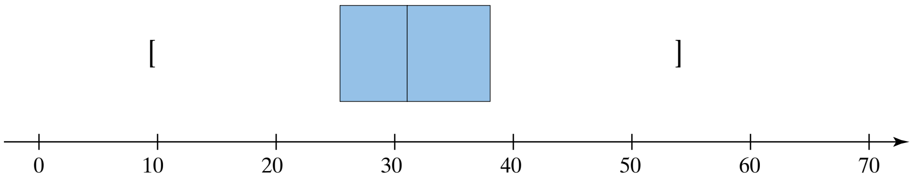
***Step 2*:** Draw a horizontal number line with a scale that will accommodate our graph. Draw vertical lines at $Q_1 = 26.06$, $M = 30.95$, and $Q_3 = 37.24$. Enclose these lines in a box. See Figure 22(a).

Figure 22(a)



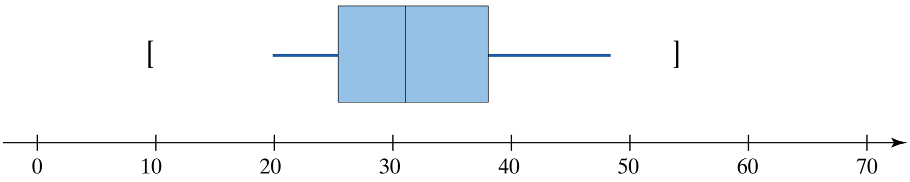
***Step 3*:** Temporarily mark the location of the lower and upper fence with brackets ([and]). See Figure 22(b).

Figure 22(b)



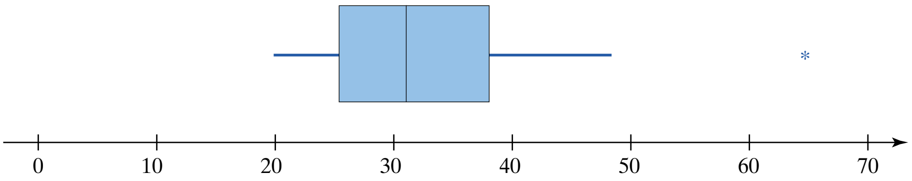
Step 4: The smallest data value that is larger than 9.29 (the lower fence) is 19.95. The largest data value that is smaller than 54.01 (the upper fence) is 49.17. We draw horizontal lines from Q_1 to 19.95 and from Q_3 to 49.17. See Figure 22(c).

Figure 22(c)



Step 5: Plot any values less than 9.29 (the lower fence) or greater than 54.01 (the upper fence) using an asterisk (*). These values are outliers. So, 64.63 is an outlier. Remove the brackets from the graph. See Figure 22(d).

Figure 22(d)



We can describe the shape of the distribution using the boxplot.

Distribution Shape Based on the Boxplot

1. If the median is near the center of the box and each horizontal line is of approximately equal length, the distribution is roughly symmetric.
2. If the median is to the left of the center of the box or the right line is substantially longer than the left line, the distribution is skewed right.
3. If the median is to the right of the center of the box or the left line is substantially longer than the right line, the distribution is skewed left.

CAUTION
Identifying the shape of a distribution from a boxplot (or from a histogram, for that matter) is subjective. When identifying the shape of a distribution from a graph, be sure to support your opinion.

Figure 23 on page 163 provides examples of boxplots that are (a) symmetric, (b) skewed right, and (c) skewed left, along with the corresponding histograms.

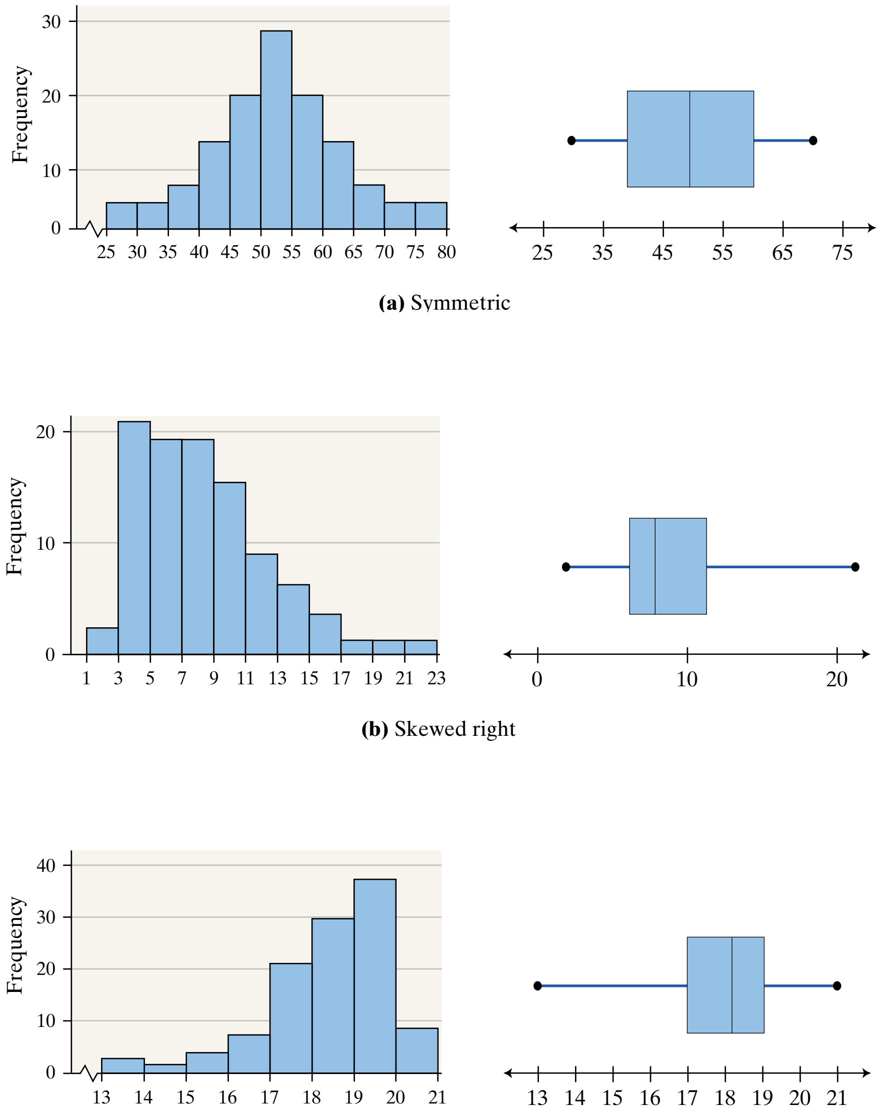
The boxplot in Figure 22(d) suggests that the distribution is skewed right, since the right line is longer than the left and the median is to the left of the center of the box.

We stated that the interquartile range (IQR) is a measure of dispersion (just like the standard deviation is a measure of dispersion). If the median is used as the measure of central tendency, then the IQR should be used as the measure of spread. Remember, we use the median to describe the "center" of a set of data when the shape of the distribution is skewed because the mean is distorted by skewness. Likewise, the standard deviation is distorted by skewness. Because the IQR is resistant to extreme values, it is a better measure of spread for skewed distributions.

Summary: Which Measures to Report

Shape of Distribution	Measure of Central Tendency	Measure of Dispersion
Symmetric	Mean	Standard deviation
Skewed left or skewed right	Median	Interquartile range

Figure 23



(a) Symmetric

(b) Skewed right

(c) Skewed left

Historical Note

John Tukey was born on July 16, 1915, in New Bedford, Massachusetts. His parents graduated numbers 1 and 2 from Bates College and were elected “the couple most likely to give birth to a genius.” In 1936, Tukey graduated from Brown University with an undergraduate degree in chemistry. He went on to earn a master’s degree in chemistry at Brown. In 1939, Tukey earned his doctorate in mathematics from Princeton. He remained at Princeton and in 1965 became the founding chair of the Department of Statistics. Among his many accomplishments, Tukey is credited with coining the terms *software* and *bit*. In the early 1970s, he discussed the negative effects of aerosol cans on the ozone layer. Tukey recommended that the 1990 Census be adjusted by means of statistical formulas. John Tukey died in New Brunswick, New Jersey, on July 26, 2000.

Now Work Problem 11.

EXAMPLE 4 Comparing Two Distributions by Using Boxplots

Problem: In the Spacelab Life Sciences 2, 14 male rats were sent to space. Upon their return, the red blood cell mass (in milliliters) of the rats was determined. A control group of 14 male rats was held under the same conditions (except for spaceflight) as the space rats, and their red blood cell mass was also determined when the space rats returned. The project was led by Paul X. Callahan. The data in Table 19 were obtained. Construct boxplots for red blood cell mass for the flight group and control group. Does it appear that the flight to space affected the red blood cell mass of the rats?

Table 19

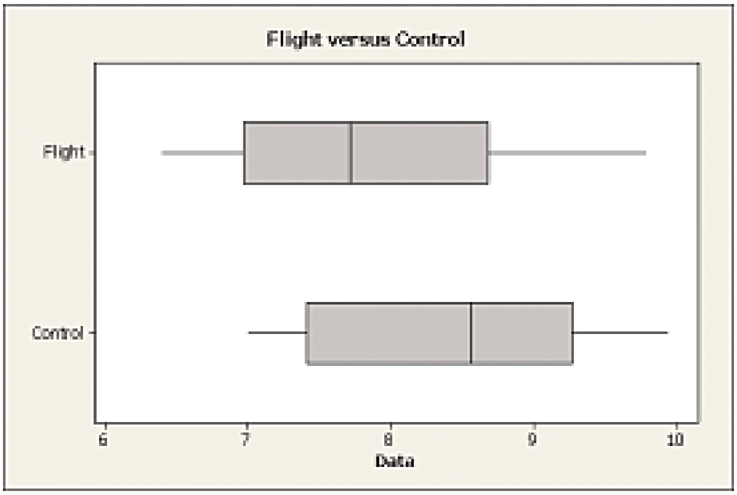
Flight		Control	
8.59	8.64	8.65	6.99
6.87	7.89	7.62	7.44
7.00	8.80	7.33	8.58
6.39	7.54	7.14	9.14
7.43	7.21	8.40	9.66
9.79	6.85	8.55	8.70
9.30	8.03	9.88	9.94

Source: NASA Life Sciences Data Archive

Approach: When comparing two data sets, we draw the boxplots on the same horizontal number line to make the comparison easy. Graphing calculators with advanced statistical features, as well as statistical spreadsheets such as MINITAB and Excel, have the ability to draw boxplots. We will use MINITAB to draw the boxplots. The steps for drawing boxplots using a TI-83 or TI-84 Plus graphing calculator, MINITAB, or Excel are given in the Technology Step by Step on page 168.

Solution: Figure 24 shows the boxplots drawn in MINITAB. From the boxplots, it appears that the spaceflight has reduced the red blood cell mass of the rats.

Figure 24



Now Work Problem 15.

MAKING AN INFORMED DECISION

What Car Should I Buy?

Suppose you are in the market to purchase a used car. To make an informed decision regarding your purchase, you would like to collect as much information as possible. Among the information you might consider are the typical price of the car, the typical number of miles the car should have and its crash test results, insurance costs, and expected repair costs.

1. Make a list of at least three cars that you would consider purchasing. To be fair, the cars should be in the same class (such as compact, midsize, and so on). They should also be of the same age.
2. Collect information regarding the three cars in your list by finding at least eight cars of each type that are for sale. Obtain such information as the asking price and the number of miles the car has. Sources of data include your local newspaper, classified ads, and car Web sites (such as www.cars.com and www.vehix.com). Compute summary statistics for asking price, number of miles, and other variables of interest. Using the same scale, draw side-by-side boxplots of each variable considered.
3. Go to the Insurance Institute for Highway Safety Web site (www.hwysafety.org). Select the Vehicle Ratings link. Choose the make and model for each car you are considering. Obtain information regarding crash testing for each car under consideration. Compare cars in the same class. How does each car compare? Is one car you are considering substantially safer than the others? What about repair costs? Compute summary statistics for crash tests and repair costs.
4. Obtain information about insurance costs. Contact various insurance companies to determine the cost of insuring the cars you are considering. Compute summary statistics for insurance costs and draw boxplots.
5. Write a report supporting your conclusion regarding which car you would purchase.

3.5 ASSESS YOUR UNDERSTANDING

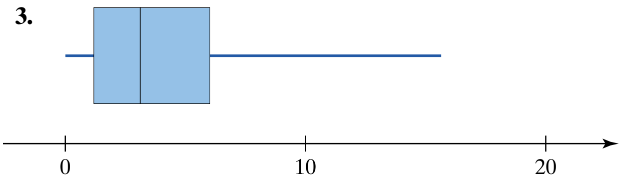
Concepts and Vocabulary

1. Explain the circumstances under which the median and interquartile range would be better measures of central tendency and dispersion than the mean and standard deviation.
2. In a boxplot, if the median is to the left of the center of the box or the right line is substantially longer than the left line, the distribution is skewed _____.

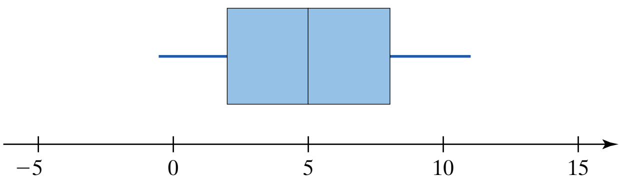
Skill Building

In Problems 3 and 4, (a) identify the shape of the distribution, and (b) determine the five-number summary. Assume that each number in the five-number summary is an integer.

3.



4.



Applying the Concepts

In Problems 5–10, find the five-number summary, and construct a boxplot for the data in the indicated problem. Comment on the shape of the distribution.

5. Age at Inauguration The following data represent the age of U.S. presidents on their respective inauguration days.

57	61	57	57	58	57	61
54	68	51	49	64	50	48
65	52	56	46	54	49	50
47	55	55	54	42	51	56
55	51	54	51	60	62	43
55	56	61	52	69	64	46
54						

6. Grams of Fat in a McDonald's Breakfast The following data represent the number of grams of fat in breakfast meals offered at McDonald's.

12	23	28	2	28	33
31	11	23	40	35	1
23	33	23	16	11	8
8	17	16	15		

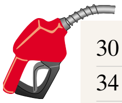
Source: McDonald's Corporation, *A Full Serving of Nutrition Facts*, April 2003

7. Super Bowl Point Spreads The following data represent the number of points by which the winning team won Super Bowls I to XXXIX.

25	19	9	16	3	21	7	17
10	4	18	17	4	12	17	5
10	29	22	36	19	32	4	45
1	13	35	17	23	10	14	7
15	7	27	3	27	3	3	

Source: superbowl.com

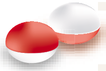
8. Miles per Gallon The following data represent the number of miles per gallon achieved on the highway for compact cars for model year 2005.



30	29	30	21	18	29	27	30	29
34	34	30	28	30	20	32	28	32
34	35	26	26	31	25	35	32	25
19	26	19	24	22	24	19	31	26
34	32	34	25	34	34	32	29	25
31	29	30	30	34	32	29	38	39
46	31	31	30	27	29	26	29	24

Source: U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, U.S. Environmental Protection Agency, *Fuel Economy Guide, Model Year 2005* (www.fueleconomy.gov)

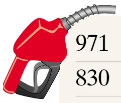
9. Got a Headache? Drugs are made of active ingredients and inactive ingredients. The Food and Drug Administration states that a drug should have the same amount of active ingredient in over-the-counter drugs. The following data represent the weight (in grams) of a random sample of 25 Tylenol tablets. What do you think is the source of the variability in weight?



0.608	0.601	0.606	0.602	0.611
0.608	0.610	0.610	0.607	0.600
0.608	0.608	0.605	0.609	0.605
0.610	0.607	0.611	0.608	0.610
0.612	0.598	0.600	0.605	0.603

Source: Kelly Roe, student at Joliet Junior College

10. Gasoline Expenditures The following data represent the mean gasoline expenditures per person for each state and the District of Columbia. Wyoming has the highest mean expenditure. What might explain this? New York has the lowest mean expenditure. What might explain this?



971	787	713	704	688	675	660	643	581
830	741	711	698	684	672	654	618	480
821	740	711	698	683	669	653	616	421
816	737	707	692	682	667	649	611	
802	726	707	692	679	666	646	598	
791	715	706	688	678	664	645	583	

Source: Energy Information Administration

11. Serum HDL NW Dr. Paul Oswiecmiski randomly selects 40 of his 20- to 29-year-old patients and obtains the following data regarding their serum HDL cholesterol.

70	56	48	48	53	52	66	48
36	49	28	35	58	62	45	60
38	73	45	51	56	51	46	39
56	32	44	60	51	44	63	50
46	69	53	70	33	54	55	52

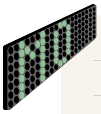
(a) Compute the five-number summary.

(b) Draw a boxplot of the data.

(c) Determine the shape of the distribution from the boxplot. Refer to the histogram drawn in Problem 31 in Section 2.2 to test your answer.

(d) Which measures of central tendency and dispersion should be reported for these data?

12. Volume of Altria Group Stock The volume of a stock is the number of shares traded on a given day. The following data, given in millions so that 3.78 represents 3,780,000 shares traded, represent the volume of Altria Group stock traded for a random sample 35 trading days in 2004.



3.78	8.74	4.35	5.02	8.40
6.06	5.75	5.34	6.92	6.23
5.32	3.25	6.57	7.57	6.07
3.04	5.64	5.00	7.16	4.88
10.32	3.38	7.25	6.52	4.43
3.38	5.53	4.74	9.70	3.56
10.96	4.50	7.97	3.01	5.58

Source: yahoo.finance.com

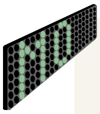
(a) Compute the five-number summary.

(b) Draw a boxplot of the data.

(c) Determine the shape of the distribution from the boxplot. Refer to the histogram drawn in Problem 32 in Section 2.2 to test your answer.

(d) Which measures of central tendency and dispersion should be reported for these data?

13. Dividend Yield A dividend is a payment from a publicly traded company to its shareholders. The dividend yield of a stock is determined by dividing the annual dividend of a stock by its price. The following data represent the dividend yields (in percent) of a random sample of 28 publicly traded stocks with a value of at least $5 billion.



1.7	0	1.15	0.62	1.06	2.45	2.38
2.83	2.16	1.05	1.22	1.68	0.89	0
2.59	0	1.7	0.64	0.67	2.07	0.94
2.04	0	0	1.35	0	0	0.41

Source: Yahoo! Finance

(a) Compute the five-number summary.

(b) Draw a boxplot of the data.

(c) Determine the shape of the distribution from the boxplot. Refer to the histogram drawn in Problem 33 in Section 2.2 to test your answer.

(d) Which measures of central tendency and dispersion should be reported for these data?

14. Violent Crimes Violent crimes include murder, forcible rape, robbery, and aggravated assault. The following data represent the violent crime rate (per 100,000 population) by state and the District of Columbia in 2002.

444	563	553	424	593	352
311	599	1633	770	459	262
255	621	357	286	377	279
662	108	770	484	540	268
343	539	352	314	638	161
375	740	496	470	78	351
503	292	402	285	822	177
717	579	237	107	291	345
234	225	274			

Source: U.S. Federal Bureau of Investigation

(a) Compute the five-number summary.

(b) Draw a boxplot of the data.

(c) Determine the shape of the distribution from the boxplot. Refer to the histogram drawn in Problem 34 in Section 2.2 to test your answer.

(d) Which measures of central tendency and dispersion should be reported for this data?

In Problems 15–17, compare the data sets by determining the five-number summary and constructing boxplots on the same scale.

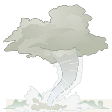
15. Chips per Cookie The data to the right represent the number of chips per cookie in a random sample of Keebler Chips Deluxe Chocolate Chip Cookies and the number of chips per cookie in a store brand's chocolate chip cookies. Does there appear to be a difference in the number of chips per cookie? Does one brand have a more consistent number of chips per cookie?

NW

Keebler			Store Brand		
32	23	28	21	23	24
28	28	29	24	25	27
25	20	25	26	26	21
22	21	24	18	16	24
21	24	21	21	30	17
26	28	24	23	28	31
33	20	31	27	33	29

Source: Trina McNamara, student at Joliet Junior College

16. Tornades The following data give the number of tornadoes in Oklahoma, Kansas, and Nebraska for the years 1990 to 2004. Which state appears to have the highest number of tornadoes per year?



Year:	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**	**2004**
Oklahoma:	30	73	64	64	40	79	47	55	83	145	44	61	18	78	62
Kansas:	88	116	92	113	42	73	68	62	71	64	59	101	95	91	122
Nebraska:	88	63	74	69	55	26	60	30	65	102	61	62	28	81	110

Source: National Oceanic and Atmospheric Administration

17. Home-Run Distances During the 1998 major league baseball season, Mark McGwire of the St. Louis Cardinals and Sammy Sosa of the Chicago Cubs thrilled fans across the country in a race to set the record for the most home runs hit in a season. Sosa ended the season with 66 home runs, and McGwire set the record with 70 home runs. Only 3 years later in 2001, Barry Bonds of the San Francisco Giants broke McGwire's record by hitting 73 home runs. The following data represent the distances of each player's home runs in his record-setting season. Which player appears to have the longest distances? Which player appears to have the most consistent distances?

Mark McGwire						
360	370	370	430	420	340	460
410	440	410	380	360	350	527
380	550	478	420	390	420	425
370	480	390	430	388	423	410
360	410	450	350	450	430	461
430	470	440	400	390	510	430
450	452	420	380	470	398	409
385	369	460	390	510	500	450
470	430	458	380	430	341	385
410	420	380	400	440	377	370

Sammy Sosa						
371	350	430	420	430	434	370
420	440	410	420	460	400	430
410	370	370	410	380	340	350
420	410	415	430	380	380	366
500	380	390	400	364	430	450
440	365	420	350	420	400	380
380	400	370	420	360	368	430
433	388	440	414	482	364	370
400	405	433	390	480	480	434
344	410	420				

Barry Bonds						
420	417	440	410	390	417	420
410	380	430	370	420	400	360
410	420	391	416	440	410	415
436	430	410	400	390	420	410
420	410	410	450	320	430	380
375	375	347	380	429	320	360
375	370	440	400	405	430	350
396	410	380	430	415	380	375
400	435	420	420	488	361	394
410	411	365	360	440	435	454
442	404	385				

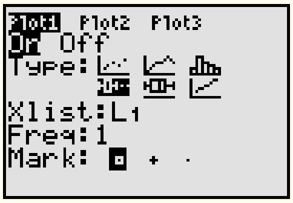
Technology Step by Step — Drawing Boxplots Using Technology

TI-83/84 Plus

***Step 1*:** Enter the raw data into L1.

***Step 2*:** Press 2nd Y= and select 1:Plot 1.

***Step 3*:** Turn the plots ON. Use the cursor to highlight the modified boxplot icon. Your screen should look as follows:



```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Freq:1
Mark: □ + ·
```

***Step 4*:** Press ZOOM and select 9: ZoomStat.

MINITAB

***Step 1*:** Enter the raw data into column C1.

***Step 2*:** Select the **Graph** menu and highlight **Boxplot . . .**

***Step 3*:** For a single boxplot, select One Y, simple. For two or more boxplots, select Multiple Y's, simple.

***Step 4*:** Select the data to be graphed. If you want the boxplot to be horizontal rather than vertical, select the Scale button, then transpose value and category scales. Click OK.

Excel

***Step 1*:** Start the PHStat Add-in.

***Step 2*:** Enter the raw data into column A.

***Step 3*:** Select the **PHStat** menu and highlight **Box-and-Whisker Plot** . . . With the cursor in the Data Variable Cell Range cell, highlight the data in column A.

***Step 4*:** Click OK.

CHAPTER 3 Review

Summary

This chapter concentrated on describing distributions numerically. Measures of central tendency are used to indicate the typical value in a distribution. Three measures of central tendency were discussed. The mean measures the center of gravity of the distribution. The median separates the bottom 50% of the data from the top 50%. Both measures require that the data be quantitative. The mode measures the most frequent observation. The data can be either quantitative or qualitative to compute the mode. The median is resistant to extreme values, while the mean is not. A comparison between the median and mean can help determine the shape of the distribution.

Measures of dispersion describe the spread in the data. The range is the difference between the highest and lowest data value. The variance measures the average squared deviation about the mean. The standard deviation is the square root of the variance. The mean and standard deviation are used in many types of statistical inference.

The mean, median, and mode can be approximated from grouped data. The variance and standard deviation can also be approximated from grouped data.

We can determine the relative position of an observation in a data set using z-scores and percentiles. z-scores denote how many standard deviations an observation is from the mean. Percentiles determine the percent of observations that lie above and below an observation. The upper and lower fences can be used to identify potential outliers. Any potential outlier must be investigated to determine whether it was the result of a data entry error, of some other error in the data collection process, or of an unusual value in the data set.

The interquartile range is also a measure of dispersion. The five-number summary provides an idea about the center and spread of a data set, through the median and the interquartile range. The length of the tails in the distribution can be determined from the smallest and largest data values. The five-number summary is used to construct boxplots. Boxplots can be used to describe the shape of the distribution.

Formulas

Population Mean

$$\mu = \frac{\sum x_i}{N}$$

Sample Mean

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

Population Variance

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}{N}$$

Sample Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n - 1}$$

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample Standard Deviation

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

Range = Largest Data Value − Smallest Data Value

Weighted Mean

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Population Mean from Grouped Data

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$

Sample Mean from Grouped Data

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

Population Variance from Grouped Data

$$\sigma^2 = \frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}$$

Sample Variance from Grouped Data

$$s^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{\left(\sum f_i\right) - 1}$$

Population *z*-Score

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

Sample *z*-Score

$$z = \frac{x - \bar{x}}{s}$$

Percentile of *x* $= \dfrac{\text{number of data values less than } x}{n} \cdot 100$

Interquartile Range

$$\text{IQR} = Q_3 - Q_1$$

Lower and Upper Fences

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR})$$

$$\text{Upper Fence} = Q_3 + 1.5(\text{IQR})$$

Vocabulary

Parameter (p. 107)
Statistic (p. 107)
Arithmetic mean (p. 107)
Median (p. 110)
Mode (p. 111)
Bimodal (p. 112)
Multimodal (p. 112)
Resistant (p. 113)
Range (p. 124)
Deviation about the mean (p. 125)
Population variance (p. 125)
Sample variance (p. 127)
Biased (p. 127)
Degrees of freedom (p. 128)
Population standard deviation (p. 129)
Sample standard deviation (p. 129)
Class midpoint (p. 142)
Weighted mean (p. 144)
Z-score (p. 150)
*k*th percentile (p. 151)
Quartiles (p. 153)
Outlier (p. 155)
Interquartile range (p. 155)
Fences (p. 155)
Exploratory data analysis (p. 159)
Five-number summary (p. 160)
Boxplot (p. 161)

Objectives

Section	You should be able to . . .	Example	Review Exercises
3.1	1 Determine the arithmetic mean of a variable from raw data (p. 107)	1, 7	1(a)–8(a) 1(a)–8(a)
	2 Determine the median of a variable from raw data (p. 110)	2, 3, 7	5(a) and 6(a)
	3 Determine the mode of a variable from raw data (p. 111)	4, 5, 6, 7	15(a)–(c), 16(a)–(c)
	4 Use the mean and median to help identify the shape of a distribution (p. 113)	8, 9	
3.2	1 Compute the range of a variable from raw data (p. 124)	2	1(b)–8(b)
	2 Compute the variance of a variable from raw data (p. 125)	3, 4, 6	1(b)–8(b), 15(d), 16(d)
	3 Compute the standard deviation of a variable from raw data (p. 129)	5, 6, 7	1(b)–8(b), 15(d), 16(d)
	4 Use the Empirical Rule to describe data that are bell shaped (p. 131)	8	9(a)–(d), 10 (a)–(c)
	5 Use Chebyshev's Inequality to describe any set of data (p. 132)	9	9(e) and (f), 10(d) and (e)
3.3	1 Approximate the mean of a variable from grouped data (p. 142)	1, 4	11(a), 12(a)
	2 Compute the weighted mean (p. 144)	2	13, 14
	3 Approximate the variance and standard deviation of a variable from grouped data (p. 144)	3, 4	11(b), 12(b)
3.4	1 Determine and interpret *z*-scores (p. 149)	1	19
	2 Determine and interpret percentiles (p. 151)	2–4	17, 18
	3 Determine and interpret quartiles (p. 153)	5, 6	15(e), 16(e)
	4 Check a set of data for outliers (p. 155)	7	15(h), 16(h)
3.5	1 Compute the five-number summary (p. 159)	1, 2	15(e), 16(e)
	2 Draw and interpret boxplots (p. 161)	3, 4	15(f), 16(f), 20

3 Review Exercises

1. **Muzzle Velocity** The following data represent the muzzle velocity (in meters per second) of rounds fired from a 155 mm gun.

793.8	793.1	792.4	794.0	791.4
792.4	791.7	792.3	789.6	794.4

Source: Christenson, Ronald, and Blackwood, Larry; "Tests for Precision and Accuracy of Multiple Measuring Devices." *Technometrics*, Nov. 93, Vol. 35, Issue 4, pp. 411–421.

(a) Compute the sample mean and median muzzle velocity.
(b) Compute the range, sample variance, and sample standard deviation.

2. **Pulse Rates** The following data represent the pulse rate of eight randomly selected females after stepping up and down on a 6-inch platform for 3 minutes. Pulse is measured in beats per minute.

136	169	120	128	129
143	115	146	96	86

Source: Michael McCraith, Joliet Junior College

(a) Compute the sample mean and median pulse.
(b) Compute the range, sample variance, and sample standard deviation.

3. **Price of Chevy Cavaliers** The following data represent the sales price in dollars for nine two-year-old Chevrolet Cavaliers in the Los Angeles area.

14,050	13,999	12,999	10,995	9,980
8,998	7,889	7,200	5,500	

Source: cars.com

(a) Compute the sample mean and median price.

(b) Compute the range and sample standard deviation.

(c) Redo (a) and (b) if the data value 14,050 was incorrectly entered as 41,050. How does this change affect the mean? The median? The range? The standard deviation? Which of these values is resistant?

4. **Home Sales** The following data represent the closing prices (in U.S. dollars) of 15 randomly selected homes sold in Joliet, Illinois, in December 2004.

138,820	140,794	136,833	157,216
169,541	153,146	115,000	149,380
135,512	99,000	124,757	136,529
149,143	136,924	128,429	

Source: Transamerica Intellitech

(a) Compute the sample mean and median sale price.

(b) Compute the range and sample standard deviation.

5. **Chief Justices** The following data represent the ages of chief justices of the U.S. Supreme Court when they were appointed.

Justice	Age
John Jay	44
John Rutledge	56
Oliver Ellsworth	51
John Marshall	46
Roger B. Taney	59
Salmon P. Chase	56
Morrison R. Waite	58
Melville W. Fuller	55
Edward D. White	65
William H. Taft	64
Charles E. Hughes	68
Harlan F. Stone	69
Frederick M. Vinson	56
Earl Warren	62
Warren E. Burger	62
William H. Rehnquist	62
John G. Roberts	50

Source: *Information Please Almanac*

(a) Compute the population mean, median, and mode ages.

(b) Compute the range and population standard deviation ages.

(c) Obtain two simple random samples of size 4, and compute the sample mean and sample standard deviation ages.

6. **National League Home Runs** The following data represent the number of home runs hit by all teams in the National League in 2004.

Team	Home Runs	Team	Home Runs
1. St. Louis Cardinals	214	9. Los Angeles Dodgers	203
2. San Francisco Giants	183	10. Cincinnati Reds	194
3. Philadelphia Phillies	215	11. Florida Marlins	148
4. Colorado Rockies	202	12. New York Mets	185
5. Atlanta Braves	178	13. Pittsburgh Pirates	142
6. Houston Astros	187	14. Montreal Expos	151
7. Chicago Cubs	235	15. Milwaukee Brewers	135
8. San Diego Padres	139	16. Arizona Diamondbacks	135

Source: Major League Baseball

(a) Compute the population mean, median, and mode for number of home runs.

(b) Compute the range and population standard deviation for number of home runs.

(c) Obtain two simple random samples of size 3, and compute the sample mean and sample standard deviation for number of home runs.

(d) If a sports reporter stated that the average number of home runs hit by teams in the National League in 2004 was 135, is he lying? Is he being deceptive?

7. **Family Size** A random sample of 36 married couples who had been married 7 years were asked the number of children they had. The results of the survey follow:

0	0	3	1	2	3
3	4	3	3	0	3
1	2	1	3	0	3
4	2	3	2	2	4
2	1	3	4	1	3
0	3	3	3	2	1

(a) Compute the sample mean and the median number of children.

(b) Compute the range and the sample standard deviation number of children.

8. **Waiting in Line** The following data represent the number of cars that arrived at a McDonald's drive-through between 11:50 A.M. and 12:00 noon each Wednesday for the past 30 weeks:

1	3	2	8	6
6	6	3	3	1
5	6	3	3	1
4	9	5	3	5
2	6	7	5	8
7	8	3	2	3

(a) Compute the sample mean and the median number of cars.

(b) Compute the range and the sample standard deviation number of cars.

9. **Chebyshev's Inequality and the Empirical Rule** Suppose that a random sample of 200 lightbulbs has a mean life of 600 hours and a standard deviation of 53 hours.

(a) A histogram of the data indicates the sample data follow a bell-shaped distribution. According to the Empirical Rule, 99.7% of lightbulbs have lifetimes between _____ and _____ hours.

(b) Assuming the data are bell shaped, determine the percentage of lightbulbs that will have a life between 494 and 706 hours.

(c) Assuming the data are bell shaped, what percentage of lightbulbs will last between 547 and 706 hours?

(d) If the company that manufactures the lightbulb guarantees to replace any bulb that does not last at least 441 hours, what percentage of lightbulbs can the firm expect to have to replace, according to the Empirical Rule?

(e) Use Chebyshev's Inequality to determine the minimum percentage of lightbulbs with a life within 2.5 standard deviations of the mean.

(f) Use Chebyshev's Inequality to determine the minimum percentage of lightbulbs with a life between 494 and 706 hours.

10. **Chebyshev's Inequality and the Empirical Rule** In a random sample of 250 toner cartridges, the mean number of pages a toner cartridge can print is 4302 and the standard deviation is 340.

(a) Suppose a histogram of the data indicates that the sample data follow a bell-shaped distribution. According to the Empirical Rule, 99.7% of toner cartridges will print between _____ and _____ pages.

(b) Assuming that the distribution of the data is bell shaped, determine the percentage of toner cartridges whose print total is between 3622 and 4982 pages.

(c) If the company that manufactures the toner cartridges guarantees to replace any cartridge that does not print at least 3622 pages, what percent of cartridges can the firm expect to be responsible for replacing, according to the Empirical Rule?

(d) Use Chebyshev's Inequality to determine the minimum percentage of toner cartridges with a page count within 1.5 standard deviations of the mean.

(e) Use Chebyshev's Inequality to determine the minimum percentage of toner cartridges that print between 3282 and 5322 pages.

11. **Vehicle Fatalities** The frequency distribution listed in the table represents the number of drivers in fatal crashes in 2003, by age, for males 20 to 84 years old.

Age	Number of Drivers	Age	Number of Drivers
20–24	6035	55–59	2355
25–29	4352	60–64	1664
30–34	4083	65–69	1173
35–39	3933	70–74	1025
40–44	4194	75–79	895
45–49	3716	80–84	744
50–54	3005		

Source: NHTSA

(a) Approximate the mean age of a male involved in a traffic fatality.

(b) Approximate the standard deviation age of a male involved in a traffic fatality.

12. **Vehicle Fatalities** The frequency distribution listed in the table represents the number of drivers in fatal crashes in 2003, by age, for females 20 to 84 years old.

Age	Number of Drivers	Age	Number of Drivers
20–24	1903	55–59	784
25–29	1415	60–64	599
30–34	1364	65–69	415
35–39	1430	70–74	482
40–44	1409	75–79	456
45–49	1242	80–84	372
50–54	1008		

Source: NHTSA

(a) Approximate the mean age of a female involved in a traffic fatality.

(b) Approximate the standard deviation age of a female involved in a traffic fatality.

(c) Compare the results to those obtained in Problem 11. How do you think an insurance company might use this information?

13. **Weighted Mean** Michael has just completed his first semester in college. He earned an A in his 5-hour calculus course, a B in his 4-hour chemistry course, an A in his 3-hour speech course, and a C in his 3-hour psychology course. Assuming an A equals 4 points, a B equals 3 points, and a C equals 2 points, determine Michael's grade-point average if grades are weighted by class hours.

14. **Weighted Mean** Yolanda wishes to develop a new type of meat loaf to sell at her restaurant. She decides to combine 2 pounds of ground sirloin (cost $2.70 per pound), 1 pound of ground turkey (cost $1.30 per pound), and $\frac{1}{2}$ pound of ground pork (cost $1.80 per pound). What is the cost per pound of the meat loaf?

15. **Mets versus Yankees** The following data represent the 2004 salaries (in dollars) of the players on the rosters of the New York Mets and the New York Yankees.

Yankees		Mets	
Player	**Salary**	**Player**	**Salary**
Bubba Crosby	301,400	Tyler Yates	300,000
Jorge De Paula	302,550	Eric Valent	302,500
Donovan Osborne	450,000	Jose Reyes	307,500
Orlando Hernandez	500,000	Dan Wheeler	311,500
Enrique Wilson	700,000	Ty Wigginton	316,000
Tony Clark	750,000	Orber Moreno	317,500
John Flaherty	775,000	Jason Phillips	318,000
Miguel Cairo	900,000	Grant Roberts	319,500
Ruben Sierra	1,000,000	Joe McEwing	500,000
Felix Heredia	1,800,000	Shane Spencer	537,500
Gabe White	1,925,000	Scott Strickland	650,000
Travis Lee	2,000,000	Scott Erickson	700,000
Jon Lieber	2,700,000	Vance Wilson	715,000
Paul Quintrill	3,000,000	Karim Garcia	800,000
Kenny Lofton	3,100,000	John Franco	1,000,000
Tom Gordon	3,500,000	Todd Zeile	1,000,000
Steve Karsay	6,000,000	Braden Looper	2,000,000
Hideki Matsui	7,000,000	Mike Stanton	3,000,000
Jose Contreras	8,500,000	David Weathers	3,933,333
Jorge Posada	9,000,000	Ricky Gutierrez	4,166,667
Javier Vazquez	9,000,000	Mike Cameron	4,333,333
Mariano Rivera	10,890,000	Steve Trachsel	5,000,000
Bernie Williams	12,357,143	Kazuo Matsui	5,033,333
Jason Giambi	12,428,571	Cliff Floyd	6,500,000
Gary Sheffield	13,000,000	Al Leiter	10,295,600
Kevin Brown	15,714,286	Tom Glavine	10,765,608
Mike Mussina	16,000,000	Mike Piazza	16,071,429
Derek Jeter	18,600,000	Mo Vaughn	17,166,667
Alex Rodriguez	22,000,000		

Source: USATODAY.com

(a) Compute the population mean salary for each team.
(b) Compute the median salary for each team.
(c) Given the results of (a) and (b), decide whether the distributions are symmetric, skewed right, or skewed left.
(d) Compute the population standard deviation salary for each team. Which team has more dispersion in its salaries?
(e) Compute the five-number summary for each team.
(f) On the same graph, draw boxplots for the two teams. Annotate the graph with some general remarks comparing the team salaries.
(g) Describe the shape of the distribution of each team, as illustrated by the boxplots. Does this confirm the result obtained in (c)?
(h) Which measure of central tendency is the better measure of central tendency? Why?

16. **Bearing Failures** An engineer is studying bearing failures for two different materials in aircraft gas turbine engines. The following data are failure times (in millions of cycles) for samples of the two material types.

Material A	3.17	4.31	4.52	4.66	5.69	5.88	6.91	8.01	8.97	11.92
Material B	5.78	6.71	6.84	7.23	8.20	9.65	13.44	14.71	16.39	24.37

(a) Compute the sample mean of the failure time for each material.
(b) Compute the median failure time for each material.
(c) Given the results of parts (a) and (b), decide whether the distributions are symmetric, skewed right, or skewed left.
(d) Compute the sample standard deviation of the failure time for each material. Which material has its failure times more dispersed?
(e) Compute the five-number summary for each material.
(f) On the same graph, draw boxplots for the two materials. Annotate the graph with some general remarks comparing the failure times.
(g) Describe the shape of the distribution of each material, as illustrated by the boxplots. Does this confirm the result obtained in part (c)?

17. **NASCAR Earnings** The following data represent the total earnings (in dollars) of drivers in the 2004 Nextel Cup Series.

1.	$51,505	23.	227,779	45.	1,095,040	67.	3,695,070
2.	53,465	24.	236,315	46.	1,133,620	68.	3,717,100
3.	53,765	25.	251,813	47.	1,217,520	69.	3,745,240
4.	56,565	26.	252,440	48.	1,259,210	70.	3,872,410
5.	57,450	27.	284,405	49.	1,275,530	71.	3,892,570
6.	57,590	28.	293,704	50.	1,333,520	72.	3,948,500
7.	58,925	29.	303,159	51.	1,349,620	73.	4,025,550
8.	65,175	30.	330,385	52.	1,410,570	74.	4,117,750
9.	70,550	31.	338,332	53.	1,461,640	75.	4,200,330
10.	101,260	32.	341,878	54.	1,985,120	76.	4,245,690
11.	107,090	33.	342,337	55.	2,337,420	77.	4,447,300
12.	111,250	34.	364,460	56.	2,471,940	78.	4,539,330
13.	116,150	35.	366,155	57.	2,666,590	79.	4,570,540
14.	116,359	36.	371,479	58.	2,780,130	80.	4,739,010
15.	116,369	37.	394,489	59.	2,929,400	81.	4,759,020
16.	124,312	38.	399,093	60.	3,044,900	82.	5,152,670
17.	139,614	39.	403,674	61.	3,250,320	83.	5,158,360
18.	144,040	40.	426,994	62.	3,443,350	84.	5,692,620
19.	154,100	41.	567,900	63.	3,483,440	85.	6,221,710
20.	160,261	42.	624,850	64.	3,583,440	86.	6,223,890
21.	171,475	43.	752,386	65.	3,675,880	87.	6,437,660
22.	186,610	44.	945,549	66.	3,676,310	88.	7,201,380

(a) Find and interpret the 40th percentile.
(b) Find and interpret the 95th percentile.
(c) Find and interpret the 10th percentile.
(d) What is the percentile rank of $4,117,750?
(e) What is the percentile rank of $116,359?

18. NASCAR Earnings Use the data in Problem 17 to answer the following:

(a) Find and interpret the 30th percentile.

(b) Find and interpret the 90th percentile.

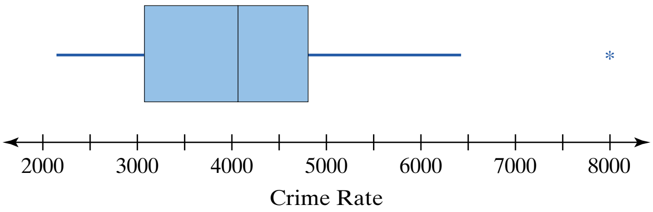
(c) Find and interpret the 5th percentile.

(d) What is the percentile rank of $1,333,520?

(e) What is the percentile rank of $139,614?

19. Weights of Males versus Females According to the National Center for Health Statistics, the mean weight of a 20- to 29-year-old female is 156.5 pounds, with a standard deviation of 51.2 pounds. The mean weight of a 20- to 29-year-old male is 183.4 pounds, with a standard deviation of 40.0 pounds. Who is relatively heavier: a 20- to 29-year-old female who weights 160 pounds or a 20- to 29-year-old male who weighs 185 pounds?

20. Crime Rate Answer the accompanying questions regarding the boxplot, which illustrates crime-rate data per 100,000 population for the 50 states and the District of Columbia in 2002. (*Source*: www.infoplease.com)



(a) Approximately what is the median crime rate in the United States?

(b) Approximately what is the 25th percentile crime rate in the United States?

(c) Are there any outliers? If so, identify their value(s).

(d) What is the lowest crime rate?

THE CHAPTER 3 CASE STUDY IS LOCATED ON THE CD THAT ACCOMPANIES THIS TEXT.