3.50 Body Temperature. A study by researchers at the University of Maryland addressed the question of whether the mean body temperature of humans is 98.6°F. The results of the study by P. Mackowiak et al. appeared in the article "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" (*Journal of the American Medical Association*, Vol. 268, pp. 1578–1580). Among other data, the researchers obtained the body temperatures of 93 healthy humans, as provided on the WeissStats CD.

*In each of Exercises **3.51–3.52**,*
a. use the technology of your choice to determine the mean and median of each of the two data sets.
b. compare the two data sets by using your results from part (a).

3.51 Treating Psychotic Illness. L. Petersen et al. evaluated the effects of integrated treatment for patients with a first episode of psychotic illness in the paper "A Randomised Multicentre Trial of Integrated Versus Standard Treatment for Patients with a First Episode of Psychotic Illness" (*British Medical Journal*, Vol. 331, (7517):602). Part of the study included a questionnaire that was designed to measure client satisfaction for both the integrated treatment and a standard treatment. The data on the WeissStats CD are based on the results of the client questionnaire.

3.52 The Etruscans. Anthropologists are still trying to unravel the mystery of the origins of the Etruscan empire, a highly advanced Italic civilization formed around the eighth century B.C. in central Italy. Were they native to the Italian peninsula or, as many aspects of their civilization suggest, did they migrate from the East by land or sea? The maximum head breadth, in millimeters, of 70 modern Italian male skulls and that of 84 preserved Etruscan male skulls were analyzed to help researchers decide whether the Etruscans were native to Italy. The resulting data can be found on the WeissStats CD. [SOURCE: N. Barnicot and D. Brothwell, "The Evaluation of Metrical Data in the Comparison of Ancient and Modern Bones." In *Medical Biology and Etruscan Origins*, G. Wolstenholme and C. O'Connor, eds., Little, Brown & Co., 1959]

Extending the Concepts and Skills

3.53 Food Choice. As you discovered earlier, *ordinal data* are data about order or rank given on a scale such as 1, 2, 3, . . . or A, B, C, ... Most statisticians recommend using the median to indicate the center of an ordinal data set, but some researchers also use the mean. In the paper "Measurement of Ethical Food Choice Motives" (*Appetite*, Vol. 34, pp. 55–59), research psychologists M. Lindeman and M. Väänänen of the University of Helsinki published a study on the factors that most influence people's choice of food. One of the questions asked of the participants was how important, on a scale of 1 to 4 (1 = not at all important, 4 = very important), is ecological welfare in food choice motive, where ecological welfare includes animal welfare and environmental protection. Here are the ratings given by 14 of the participants.

2	4	1	2	4	3	3
2	2	1	2	4	2	3

a. Compute the mean of the data.
b. Compute the median of the data.
c. Decide which of the two measures of center is best.

3.54 Outliers and Trimmed Means. Some data sets contain *outliers,* observations that fall well outside the overall pattern of the data. (We discuss outliers in more detail in Section 3.3.) Suppose, for instance, that you are interested in the ability of high school algebra students to compute square roots. You decide to give a square-root exam to 10 of these students. Unfortunately, one of the students had a fight with his girlfriend and cannot concentrate—he gets a 0. The 10 scores are displayed in increasing order in the following table. The score of 0 is an outlier.

0	58	61	63	67	69	70	71	78	80

Statisticians have a systematic method for avoiding extreme observations and outliers when they calculate means. They compute *trimmed means,* in which high and low observations are deleted or "trimmed off" before the mean is calculated. For instance, to compute the 10% trimmed mean of the test-score data, we first delete both the bottom 10% and the top 10% of the ordered data, that is, 0 and 80. Then we calculate the mean of the remaining data. Thus the 10% trimmed mean of the test-score data is

$$\frac{58+61+63+67+69+70+71+78}{8} = 67.1.$$

The following table displays a set of scores for a 40-question algebra final exam.

2	15	16	16	19	21	21	25	26	27
4	15	16	17	20	21	24	25	27	28

a. Do any of the scores look like outliers?
b. Compute the usual mean of the data.
c. Compute the 5% trimmed mean of the data.
d. Compute the 10% trimmed mean of the data.
e. Compare the means you obtained in parts (b)–(d). Which of the three means provides the best measure of center for the data?

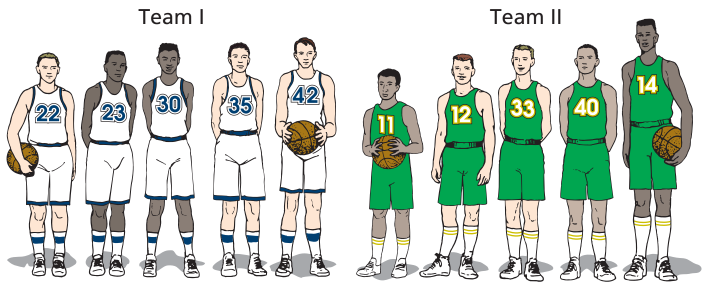
3.55 Explain the difference between the quantities $(\Sigma x_i)^2$ and Σx_i^2. Construct an example to show that, in general, those two quantities are unequal.

3.56 Explain the difference between the quantities $\Sigma x_i y_i$ and $(\Sigma x_i)(\Sigma y_i)$. Provide an example to show that, in general, those two quantities are unequal.

3.2 Measures of Variation

Up to this point, we have discussed only descriptive measures of center, specifically, the mean, median, and mode. However, two data sets can have the same mean, median, or mode and still differ in other respects. For example, consider the heights of the five starting players on each of two men's college basketball teams, as shown in Fig. 3.2.

FIGURE 3.2
Five starting players on two basketball teams



Feet and inches	6'	6'1"	6'4"	6'4"	6'6"	5'7"	6'	6'4"	6'4"	7'
Inches	72	73	76	76	78	67	72	76	76	84

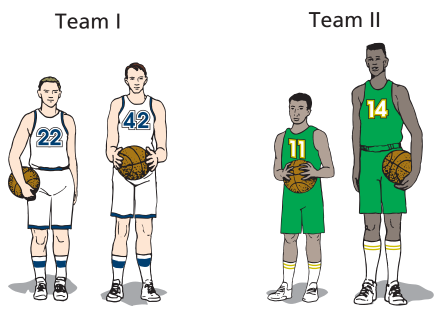
The two teams have the same mean height, 75 inches (6′ 3″); the same median height, 76 inches (6′ 4″); and the same mode, 76 inches (6′ 4″). Nonetheless, the two data sets clearly differ. In particular, the heights of the players on Team II vary much more than those on Team I. To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation, or spread, in a data set. Such descriptive measures are referred to as **measures of variation** or **measures of spread.**

Just as there are several different measures of center, there are also several different measures of variation. In this section, we examine two of the most frequently used measures of variation: the *range* and *sample standard deviation.*

The Range

The contrast between the height difference of the two teams is clear if we place the shortest player on each team next to the tallest, as in Fig. 3.3.

FIGURE 3.3
Shortest and tallest starting players on the teams



Feet and inches	6'	6'6"	5'7"	7'
Inches	72	78	67	84

The **range** of a data set is the difference between the maximum (largest) and minimum (smallest) observations. From Fig. 3.3,

$$\text{Team I: Range} = 78 - 72 = 6 \text{ inches,}$$
$$\text{Team II: Range} = 84 - 67 = 17 \text{ inches.}$$

SC Report 3.4

You try it!

Exercise 3.63(a) on page 110

Interpretation The difference between the heights of the tallest and shortest players on Team I is 6 inches, whereas that difference for Team II is 17 inches.

DEFINITION 3.5

Range of a Data Set

The **range** of a data set is given by the formula

$$\text{Range} = \text{Max} - \text{Min},$$

where Max and Min denote the maximum and minimum observations, respectively.

What Does It Mean?

The range of a data set is the difference between its largest and smallest values.

The range of a data set is easy to compute, but takes into account only the largest and smallest observations. For that reason, two other measures of variation, the *standard deviation* and the *interquartile range,* are generally favored over the range. We discuss the standard deviation in this section and consider the interquartile range in Section 3.3.

The Sample Standard Deviation

In contrast to the range, the standard deviation takes into account all the observations. It is the preferred measure of variation when the mean is used as the measure of center.

Roughly speaking, the **standard deviation** measures variation by indicating how far, on average, the observations are from the mean. For a data set with a large amount of variation, the observations will, on average, be far from the mean; so the standard deviation will be large. For a data set with a small amount of variation, the observations will, on average, be close to the mean; so the standard deviation will be small.

The formulas for the standard deviations of sample data and population data differ slightly. In this section, we concentrate on the sample standard deviation. We discuss the population standard deviation in Section 3.4.

The first step in computing a sample standard deviation is to find the **deviations from the mean,** that is, how far each observation is from the mean.

EXAMPLE 3.8 The Deviations From the Mean

Heights of Starting Players The heights, in inches, of the five starting players on Team I are 72, 73, 76, 76, and 78, as we saw in Fig. 3.2. Find the deviations from the mean.

Solution The mean height of the starting players on Team I is

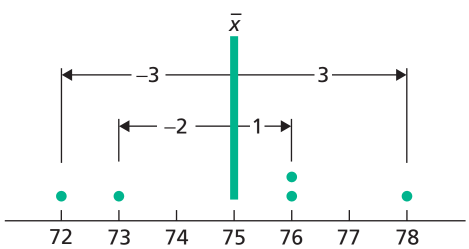
$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{72 + 73 + 76 + 76 + 78}{5} = \frac{375}{5} = 75 \text{ inches.}$$

To find the deviation from the mean for an observation x_i, we subtract the mean from it; that is, we compute $x_i - \bar{x}$. For instance, the deviation from the mean for the height of 72 inches is $x_i - \bar{x} = 72 - 75 = -3$. The deviations from the mean for all five observations are given in the second column of Table 3.6 and are represented by arrows in Fig. 3.4.

TABLE 3.6
Deviations from the mean

Height x	Deviation from mean $x - \bar{x}$
72	−3
73	−2
76	1
76	1
78	3

FIGURE 3.4
Observations (shown by dots) and deviations from the mean (shown by arrows)



The second step in computing a sample standard deviation is to obtain a measure of the total deviation from the mean for all the observations. Although the quantities $x_i - \bar{x}$ represent deviations from the mean, adding them to get a total deviation from the mean is of no value because their sum, $\Sigma(x_i - \bar{x})$, always equals zero. Summing the second column of Table 3.6 illustrates this fact for the height data of Team I.

To obtain quantities that do not sum to zero, we square the deviations from the mean. The sum of the squared deviations from the mean, $\Sigma(x_i - \bar{x})^2$, is called the **sum of squared deviations** and gives a measure of total deviation from the mean for all the observations. We show how to calculate it next.

EXAMPLE 3.9 The Sum of Squared Deviations

Heights of Starting Players Compute the sum of squared deviations for the heights of the starting players on Team I.

Solution To get Table 3.7, we added a column for $(x - \bar{x})^2$ to Table 3.6.

TABLE 3.7
Table for computing the sum of squared deviations for the heights of Team I

Height x	Deviation from mean $x - \bar{x}$	Squared deviation $(x - \bar{x})^2$
72	−3	9
73	−2	4
76	1	1
76	1	1
78	3	9
		24

From the third column of Table 3.7, $\Sigma(x_i - \bar{x})^2 = 24$. The sum of squared deviations is 24 inches^2.

The third step in computing a sample standard deviation is to take an average of the squared deviations. We do so by dividing the sum of squared deviations by $n - 1$, or 1 less than the sample size. The resulting quantity is called a **sample variance** and is denoted $\boldsymbol{s_x^2}$ or, when no confusion can arise, $\boldsymbol{s^2}$. In symbols,

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}.$$

Note: If we divided by n instead of by $n - 1$, the sample variance would be the mean of the squared deviations. Although dividing by n seems more natural, we divide by $n - 1$ for the following reason. One of the main uses of the sample variance is to estimate the population variance (defined in Section 3.4). Division by n tends to underestimate the population variance, whereas division by $n - 1$ gives, on average, the correct value.

EXAMPLE 3.10 The Sample Variance

Heights of Starting Players Determine the sample variance of the heights of the starting players on Team I.

Solution From Example 3.9, the sum of squared deviations is 24 inches^2. Because $n = 5$,

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{24}{5-1} = 6.$$

The sample variance is 6 inches^2.

As we have just seen, a sample variance is in units that are the square of the original units, the result of squaring the deviations from the mean. Because descriptive measures should be expressed in the original units, the final step in computing a sample standard deviation is to take the square root of the sample variance, which gives us the following definition.

DEFINITION 3.6

What Does It Mean?

Roughly speaking, the sample standard deviation indicates how far, on average, the observations in the sample are from the mean of the sample.

Sample Standard Deviation

For a variable x, the standard deviation of the observations for a sample is called a **sample standard deviation.** It is denoted s_x or, when no confusion will arise, simply s. We have

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}},$$

where n is the sample size and $\bar{x}$ is the sample mean.

EXAMPLE 3.11 The Sample Standard Deviation

Heights of Starting Players Determine the sample standard deviation of the heights of the starting players on Team I.

Solution From Example 3.10, the sample variance is 6 inches^2. Thus the sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{6} = 2.4 \text{ inches (rounded).}$$

Interpretation Roughly speaking, on average, the heights of the players on Team I vary from the mean height of 75 inches by about 2.4 inches.

For teaching purposes, we spread our calculations of a sample standard deviation over four separate examples. Now we summarize the procedure with three steps.

Step 1 **Calculate the sample mean, $\bar{x}$.**

Step 2 **Construct a table to obtain the sum of squared deviations, $\Sigma(x_i - \bar{x})^2$.**

Step 3 **Apply Definition 3.6 to determine the sample standard deviation, s.**

EXAMPLE 3.12 The Sample Standard Deviation

Heights of Starting Players The heights, in inches, of the five starting players on Team II are 67, 72, 76, 76, and 84. Determine the sample standard deviation of these heights.

TABLE 3.8
Table for computing the sum of squared deviations for the heights of Team II

x	$x - \bar{x}$	$(x - \bar{x})^2$
67	−8	64
72	−3	9
76	1	1
76	1	1
84	9	81
		156

Solution We apply the three-step procedure just described.

Step 1 Calculate the sample mean, $\bar{x}$.

We have

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{67 + 72 + 76 + 76 + 84}{5} = \frac{375}{5} = 75 \text{ inches.}$$

Step 2 Construct a table to obtain the sum of squared deviations, $\Sigma(x_i - \bar{x})^2$.

Table 3.8 provides columns for x, $x - \bar{x}$, and $(x - \bar{x})^2$. The third column shows that $\Sigma(x_i - \bar{x})^2 = 156$ inches^2.

Step 3 Apply Definition 3.6 to determine the sample standard deviation, s.

Because $n = 5$ and $\Sigma(x_i - \bar{x})^2 = 156$, the sample standard deviation is

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{156}{5 - 1}} = \sqrt{39} = 6.2 \text{ inches (rounded).}$$

Interpretation Roughly speaking, on average, the heights of the players on Team II vary from the mean height of 75 inches by about 6.2 inches.

You try it!

Exercise 3.63(b) on page 110

SC Report 3.5

In Examples 3.11 and 3.12, we found that the sample standard deviations of the heights of the starting players on Teams I and II are 2.4 inches and 6.2 inches, respectively. Hence Team II, which has more variation in height than Team I, also has a larger standard deviation.

KEY FACT 3.1

Variation and the Standard Deviation

The more variation that there is in a data set, the larger is its standard deviation.

APPLET Applet 3.2

Key Fact 3.1 shows that the standard deviation satisfies the basic criterion for a measure of variation; in fact, the standard deviation is the most commonly used measure of variation. However, the standard deviation does have its drawbacks. For instance, it is not resistant: its value can be strongly affected by a few extreme observations.

A Computing Formula for s

Next, we present an alternative formula for obtaining a sample standard deviation, which we call the *computing formula* for s. We call the original formula given in Definition 3.6 the *defining formula* for s.

FORMULA 3.1

Computing Formula for a Sample Standard Deviation

A sample standard deviation can be computed using the formula

$$s = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/n}{n - 1}},$$

where n is the sample size.

Note: In the numerator of the computing formula, the division of $(\Sigma x_i)^2$ by n should be performed before the subtraction from Σx_i^2. In other words, first compute $(\Sigma x_i)^2/n$ and then subtract the result from Σx_i^2.

The computing formula for s is equivalent to the defining formula—both formulas give the same answer, although differences owing to roundoff error are possible. However, the computing formula is usually faster and easier for doing calculations by hand and also reduces the chance for roundoff error.

Before illustrating the computing formula for s, let's investigate its expressions, Σx_i^2 and $(\Sigma x_i)^2$. The expression Σx_i^2 represents the sum of the squares of the data; to find it, first square each observation and then sum those squared values. The expression $(\Sigma x_i)^2$ represents the square of the sum of the data; to find it, first sum the observations and then square that sum.

EXAMPLE 3.13 Computing Formula for a Sample Standard Deviation

TABLE 3.9
Table for computation of s, using the computing formula

x	x^2
67	4,489
72	5,184
76	5,776
76	5,776
84	7,056
375	28,281

Heights of Starting Players Find the sample standard deviation of the heights for the five starting players on Team II by using the computing formula.

Solution We need the sums Σx_i^2 and $(\Sigma x_i)^2$, which Table 3.9 shows to be 375 and 28,281, respectively. Now applying Formula 3.1, we get

$$s = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2/n}{n-1}} = \sqrt{\frac{28{,}281 - (375)^2/5}{5-1}}$$
$$= \sqrt{\frac{28{,}281 - 28{,}125}{4}} = \sqrt{\frac{156}{4}} = \sqrt{39} = 6.2 \text{ inches},$$

which is the same value that we got by using the defining formula.

Rounding Basics

You try it!

Exercise 3.63(c) on page 110

Here is an important rule to remember when you use only basic calculator functions to obtain a sample standard deviation or any other descriptive measure.

Rounding Rule: Do not perform any rounding until the computation is complete; otherwise, substantial roundoff error can result.

Another common rounding rule is to round final answers that contain units to one more decimal place than the raw data. Although we usually abide by this convention, occasionally we vary from it for pedagogical reasons. In general, you should stick to this rounding rule as well.

Further Interpretation of the Standard Deviation

Again, the standard deviation is a measure of variation—the more variation there is in a data set, the larger is its standard deviation. Table 3.10 contains two data sets, each with 10 observations. Notice that Data Set II has more variation than Data Set I.

TABLE 3.10
Data sets that have different variation

Data Set I	41	44	45	47	47	48	51	53	58	66
Data Set II	20	37	48	48	49	50	53	61	64	70

TABLE 3.11
Means and standard deviations of the data sets in Table 3.10

Data Set I	Data Set II
$\bar{x} = 50.0$ $s = 7.4$	$\bar{x} = 50.0$ $s = 14.2$

We computed the sample mean and sample standard deviation of each data set and summarized the results in Table 3.11. As expected, the standard deviation of Data Set II is larger than that of Data Set I.

To enable you to compare visually the variations in the two data sets, we produced the graphs shown in Figs. 3.5 and 3.6. On each graph, we marked the observations with dots. In addition, we located the sample mean, $\bar{x} = 50$, and measured intervals equal in length to the standard deviation: 7.4 for Data Set I and 14.2 for Data Set II.

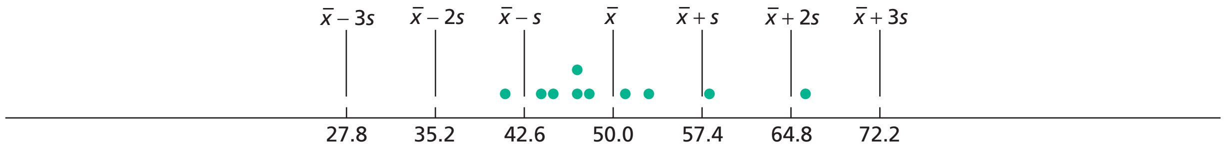


FIGURE 3.5 Data Set I; $\bar{x} = 50$, $s = 7.4$

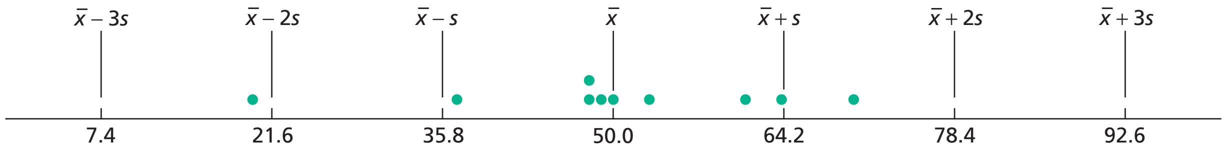


FIGURE 3.6 Data Set II; $\bar{x} = 50$, $s = 14.2$

In Fig. 3.5, note that the horizontal position labeled $\bar{x} + 2s$ represents the number that is two standard deviations to the right of the mean, which in this case is

$$\bar{x} + 2s = 50.0 + 2 \cdot 7.4 = 50.0 + 14.8 = 64.8.^{\dagger}$$

Likewise, the horizontal position labeled $\bar{x} - 3s$ represents the number that is three standard deviations to the left of the mean, which in this case is

$$\bar{x} - 3s = 50.0 - 3 \cdot 7.4 = 50.0 - 22.2 = 27.8.$$

Figure 3.6 is interpreted in a similar manner.

The graphs shown in Figs. 3.5 and 3.6 vividly illustrate that Data Set II has more variation than Data Set I. They also show that for each data set, all observations lie within a few standard deviations to either side of the mean. This result is no accident.

KEY FACT 3.2

Three-Standard-Deviations Rule

Almost all the observations in any data set lie within three standard deviations to either side of the mean.

A data set with a great deal of variation has a large standard deviation, so three standard deviations to either side of its mean will be extensive, as shown in Fig. 3.6. A data set with little variation has a small standard deviation, and hence three standard deviations to either side of its mean will be narrow, as shown in Fig. 3.5.

The three-standard-deviations rule is vague—what does "almost all" mean? It can be made more precise in several ways, two of which we now briefly describe.

We can apply **Chebychev's rule,** which is valid for all data sets and implies, in particular, that at least 89% of the observations lie within three standard deviations to either side of the mean. If the distribution of the data set is approximately bell shaped, we can apply the **empirical rule,** which implies, in particular, that roughly 99.7% of the observations lie within three standard deviations to either side of the mean. Both Chebychev's rule and the empirical rule are discussed in detail in the exercises of this section.

THE TECHNOLOGY CENTER

In Section 3.1, we showed how to use Minitab, Excel, and the TI-83/84 Plus to obtain several descriptive measures. We can apply those same programs to obtain the range and sample standard deviation.

---

† Recall that the rules for the order of arithmetic operations say to multiply and divide before adding and subtracting. So, to evaluate $a + b \cdot c$, find $b \cdot c$ first and then add the result to a. Similarly, to evaluate $a - b \cdot c$, find $b \cdot c$ first and then subtract the result from a.

EXAMPLE 3.14 Using Technology to Obtain Descriptive Measures

Heights of Starting Players The first column of Table 3.9 on page 107 gives the heights of the five starting players on Team II. Use Minitab, Excel, or the TI-83/84 Plus to find the range and sample standard deviation of those heights.

Solution We applied the descriptive-measures programs to the data, resulting in Output 3.2. Steps for generating that output are presented in Instructions 3.2.

OUTPUT 3.2 Descriptive measures for the heights of the players on Team II

MINITAB

Descriptive Statistics: HEIGHT

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
HEIGHT	5	0	75.00	2.79	6.24	67.00	69.50	76.00	80.00	84.00

EXCEL

Summary of HEIGHT

Count	5
Mean	75
Median	76
Std Dev	6.245
Variance	39
Range	17
Min	67
Max	84
IQR	7.25
25th%	70.75
75th%	78

TI-83/84 PLUS

```
1-Var Stats
 x̄=75
 Σx=375
 Σx²=28281
 Sx=6.244997998
 σx=5.585696018
↓n=5
```

```
1-Var Stats
↑n=5
 minX=67
 Q1=69.5
 Med=76
 Q3=80
 maxX=84
```

As shown in Output 3.2, the sample standard deviation of the heights for the starting players on Team II is 6.24 inches (to two decimal places). The Excel output also shows that the range of the heights is 17 inches. We can get the range from the Minitab or TI-83/84 Plus output by subtracting the minimum from the maximum.

INSTRUCTIONS 3.2 Steps for generating Output 3.2

MINITAB

1 Store the height data from Table 3.9 in a column named HEIGHT
2 Choose **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics...**
3 Specify HEIGHT in the **Variables** text box
4 Click **OK**

EXCEL

1 Store the height data from Table 3.9 in a range named HEIGHT
2 Choose **DDXL ➤ Summaries**
3 Select **Summary of One Variable** from the **Function type** drop-down box
4 Specify HEIGHT in the **Quantitative Variable** text box
5 Click **OK**

TI-83/84 PLUS

1 Store the height data from Table 3.9 in a list named HT
2 Press **STAT**
3 Arrow over to **CALC**
4 Press **1**
5 Press **2nd ➤ LIST**
6 Arrow down to HT and press **ENTER** twice

Note to Minitab users: The range is optionally available from the **Display Descriptive Statistics** dialog box by first clicking the **Statistics...** button and then checking the **Range** check box.

Exercises 3.2

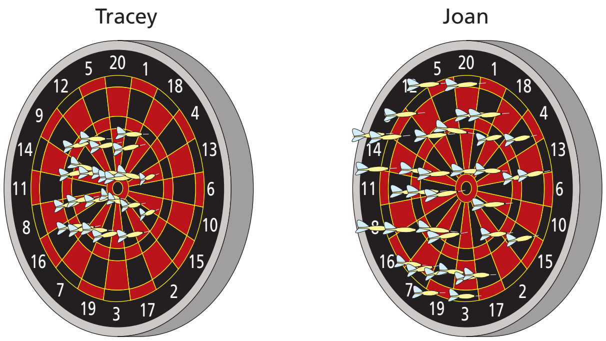
Understanding the Concepts and Skills

3.57 Explain the purpose of a measure of variation.

3.58 Why is the standard deviation preferable to the range as a measure of variation?

3.59 When you use the standard deviation as a measure of variation, what is the reference point?

3.60 Darts. The following dartboards represent darts thrown by two players, Tracey and Joan.



For the variable "distance from the center," which player's board represents data with a smaller sample standard deviation? Explain your answer.

3.61 Consider the data set 1, 2, 3, 4, 5, 6, 7, 8, 9.
a. Use the defining formula to obtain the sample standard deviation.
b. Replace the 9 in the data set by 99, and again use the defining formula to compute the sample standard deviation.
c. Compare your answers in parts (a) and (b). The lack of what property of the standard deviation accounts for its extreme sensitivity to the change of 9 to 99?

3.62 Consider the following four data sets.

Data Set I		Data Set II		Data Set III		Data Set IV	
1	5	1	9	5	5	2	4
1	8	1	9	5	5	4	4
2	8	1	9	5	5	4	4
2	9	1	9	5	5	4	10
5	9	1	9	5	5	4	10

a. Compute the mean of each data set.
b. Although the four data sets have the same means, in what respect are they quite different?
c. Which data set appears to have the least variation? the greatest variation?
d. Compute the range of each data set.
e. Use the defining formula to compute the sample standard deviation of each data set.
f. From your answers to parts (d) and (e), which measure of variation better distinguishes the spread in the four data sets: the range or the standard deviation? Explain your answer.
g. Are your answers from parts (c) and (e) consistent?

3.63 Age of U.S. Residents. The U.S. Census Bureau publishes information about ages of people in the United States in *Current Population Reports*. A sample of five U.S. residents have the following ages, in years.

21	54	9	45	51

a. Determine the range of these ages.
b. Find the sample standard deviation of these ages by using the defining formula, Definition 3.6 on page 105.
c. Find the sample standard deviation of these ages by using the computing formula, Formula 3.1 on page 106.
d. Compare your work in parts (b) and (c).

3.64 Consider the data set 3, 3, 3, 3, 3, 3.
a. Guess the value of the sample standard deviation without calculating it. Explain your reasoning.
b. Use the defining formula to calculate the sample standard deviation.
c. Complete the following statement and explain your reasoning: If all observations in a data set are equal, the sample standard deviation is ______.
d. Complete the following statement and explain your reasoning: If the sample standard deviation of a data set is 0, then....

*In Exercises **3.65–3.70**, we have provided simple data sets for you to practice the basics of finding measures of variation. For each data set, determine the*
a. range. *b. sample standard deviation.*

3.65 4, 0, 5 **3.66** 3, 5, 7

3.67 1, 2, 4, 4 **3.68** 2, 5, 0, −1

3.69 1, 9, 8, 4, 3 **3.70** 4, 2, 0, 2, 2

*In Exercises **3.71–3.78**, determine the range and sample standard deviation for each of the data sets. For the sample standard deviation, round each answer to one more decimal place than that used for the observations.*

3.71 Amphibian Embryos. In a study of the effects of radiation on amphibian embryos titled "Shedding Light on Ultraviolet Radiation and Amphibian Embryos" (*BioScience*, Vol. 53, No. 6, pp. 551–561), L. Licht recorded the time it took for a sample of seven different species of frogs' and toads' eggs to hatch. The following table shows the times to hatch, in days.

6	7	11	6	5	5	11

3.72 Hurricanes. An article by D. Schaefer et al. (*Journal of Tropical Ecology*, Vol. 16, pp. 189–207) reported on a long-term study of the effects of hurricanes on tropical streams of the Luquillo Experimental Forest in Puerto Rico. The study showed that Hurricane Hugo had a significant impact on stream water chemistry. The following table shows a sample of 10 ammonia fluxes in the first year after Hugo. Data are in kilograms per hectare per year.

96	66	147	147	175
116	57	154	88	154

3.73 Tornado Touchdowns. Each year, tornadoes that touch down are recorded by the Storm Prediction Center and published in *Monthly Tornado Statistics*. The following table gives the number of tornadoes that touched down in the United States during each month of one year. [SOURCE: National Oceanic and Atmospheric Administration.]

3	2	47	118	204	97
68	86	62	57	98	99

3.74 Technical Merit. In one Winter Olympics, Michelle Kwan competed in the Short Program ladies singles event. From nine judges, she received scores ranging from 1 (poor) to 6 (perfect). The following table provides the scores that the judges gave her on technical merit, found in an article by S. Berry (*Chance*, Vol. 15, No. 2, pp. 14–18).

5.8	5.7	5.9	5.7	5.5	5.7	5.7	5.7	5.6

3.75 Billionaires' Club. Each year, *Forbes* magazine compiles a list of the 400 richest Americans. As of September 17, 2008, the top 10 on the list are as shown in the following table.

Person	Wealth ($ billions)
William Gates III	57.0
Warren Buffett	50.0
Lawrence Ellison	27.0
Jim Walton	23.4
S. Robson Walton	23.3
Alice Walton	23.2
Christy Walton & family	23.2
Michael Bloomberg	20.0
Charles Koch	19.0
David Koch	19.0

3.76 AML and the Cost of Labor. Active Management of Labor (AML) was introduced in the 1960s to reduce the amount of time a woman spends in labor during the birth process. R. Rogers et al. conducted a study to determine whether AML also translates into a reduction in delivery cost to the patient. They reported their findings in the paper "Active Management of Labor: A Cost Analysis of a Randomized Controlled Trial" (*Western Journal of Medicine*, Vol. 172, pp. 240–243). The following table displays the costs, in dollars, of eight randomly sampled AML deliveries.

3141	2873	2116	1684
3470	1799	2539	3093

3.77 Fuel Economy. Every year, *Consumer Reports* publishes a magazine titled *New Car Ratings and Review* that looks at vehicle profiles for the year's models. It lets you see in one place how, within each category, the vehicles compare. One category of interest, especially when fuel prices are rising, is fuel economy, measured in miles per gallon (mpg). Following is a list of overall mpg for 14 different full-sized and compact pickups.

14	13	14	13	14	14	11
12	15	15	17	14	15	16

3.78 Router Horsepower. In the article "Router Roundup" (*Popular Mechanics*, Vol. 180, No. 12, pp. 104–109), T. Klenck reported on tests of seven fixed-base routers for performance, features, and handling. The following table gives the horsepower for each of the seven routers tested.

1.75	2.25	2.25	2.25	1.75	2.00	1.50

3.79 Medieval Cremation Burials. In the article "Material Culture as Memory: Combs and Cremations in Early Medieval Britain" (*Early Medieval Europe*, Vol. 12, Issue 2, pp. 89–128), H. Williams discussed the frequency of cremation burials found in 17 archaeological sites in eastern England. Here are the data.

83	64	46	48	523	35	34	265	2484
46	385	21	86	429	51	258	119	

a. Obtain the sample standard deviation of these data.
b. Do you think that, in this case, the sample standard deviation provides a good measure of variation? Explain your answer.

3.80 Monthly Motorcycle Casualties. The *Scottish Executive*, Analytical Services Division Transport Statistics, compiles data on motorcycle casualties. During one year, monthly casualties resulting from motorcycle accidents in Scotland for built-up roads and non–built-up roads were as follows.

Month	Built-up	Non built-up
January	25	16
February	38	9
March	38	26
April	56	48
May	61	73
June	52	72
July	50	91
August	90	69
September	67	71
October	51	28
November	64	19
December	40	12

a. Without doing any calculations, make an educated guess at which of the two data sets, built-up or non built-up, has the greater variation.
b. Find the range and sample standard deviation of each of the two data sets. Compare your results here to the educated guess that you made in part (a).

3.81 Daily Motorcycle Accidents. The *Scottish Executive*, Analytical Services Division Transport Statistics, compiles data on motorcycle accidents. During one year, the numbers of motorcycle accidents in Scotland were tabulated by day of the week for built-up roads and non–built-up roads and resulted in the following data.

Day	Built-up	Non built-up
Monday	88	70
Tuesday	100	58
Wednesday	76	59
Thursday	98	53
Friday	103	56
Saturday	85	94
Sunday	69	102

a. Without doing any calculations, make an educated guess at which of the two data sets, built-up or non built-up, has the greater variation.
b. Find the range and sample standard deviation of each of the two data sets. Compare your results here to the educated guess that you made in part (a).

Working with Large Data Sets

*In each of Exercises **3.82–3.90**, use the technology of your choice to determine and interpret the range and sample standard deviation for those data sets to which those concepts apply. If those concepts don't apply, explain why. Note: If an exercise contains more than one data set, perform the aforementioned tasks for each data set.*

3.82 Car Sales. The American Automobile Manufacturers Association compiles data on U.S. car sales by type of car. Results are published in the *World Almanac*. A random sample of last year's car sales yielded the car-type data on the WeissStats CD.

3.83 U.S. Hospitals. The American Hospital Association conducts annual surveys of hospitals in the United States and publishes its findings in *AHA Hospital Statistics*. Data on hospital type for U.S. registered hospitals can be found on the WeissStats CD. For convenience, we use the following abbreviations:

- NPC: Nongovernment not-for-profit community hospitals
- IOC: Investor-owned (for-profit) community hospitals
- SLC: State and local government community hospitals
- FGH: Federal government hospitals
- NFP: Nonfederal psychiatric hospitals
- NLT: Nonfederal long-term-care hospitals
- HUI: Hospital units of institutions

3.84 Marital Status and Drinking. Research by W. Clark and L. Midanik (*Alcohol Consumption and Related Problems: Alcohol and Health Monograph 1*. DHHS Pub. No. (ADM) 82–1190) examined, among other issues, alcohol consumption patterns of U.S. adults by marital status. Data for marital status and number of drinks per month, based on the researcher's survey results, are provided on the WeissStats CD.

3.85 Ballot Preferences. In Issue 338 of the *Amstat News*, then-president of the American Statistical Association, F. Scheuren reported the results of a survey on how members would prefer to receive ballots in annual elections. On the WeissStats CD, you will find data for preference and highest degree obtained for the 566 respondents.

3.86 The Great White Shark. In an article titled "Great White, Deep Trouble" (*National Geographic*, Vol. 197(4), pp. 2–29), Peter Benchley—the author of *JAWS*—discussed various aspects of the Great White Shark (*Carcharodon carcharias*). Data on the number of pups borne in a lifetime by each of 80 Great White Shark females are provided on the WeissStats CD.

3.87 Top Recording Artists. From the Recording Industry Association of America Web site, we obtained data on the number of albums sold, in millions, for the top recording artists (U.S. sales only) as of November 6, 2008. Those data are provided on the WeissStats CD.

3.88 Educational Attainment. As reported by the U.S. Census Bureau in *Current Population Reports*, the percentage of adults in each state and the District of Columbia who have completed high school is provided on the WeissStats CD.

3.89 Crime Rates. The U.S. Federal Bureau of Investigation publishes the annual crime rates for each state and the District of Columbia in the document *Crime in the United States*. Those rates, given per 1000 population, are provided on the WeissStats CD.

3.90 Body Temperature. A study by researchers at the University of Maryland addressed the question of whether the mean body temperature of humans is 98.6°F. The results of the study by P. Mackowiak et al. appeared in the article "A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich" (*Journal of the American Medical Association*, Vol. 268, pp. 1578–1580). Among other data, the researchers obtained the body temperatures of 93 healthy humans, as provided on the WeissStats CD.

*In each of Exercises **3.91–3.92**,*
a. use the technology of your choice to determine the range and sample standard deviation of each of the two data sets.
b. compare the two data sets by using your results from part (a).

3.91 Treating Psychotic Illness. L. Petersen et al. evaluated the effects of integrated treatment for patients with a first episode of psychotic illness in the paper "A Randomised Multicentre Trial of Integrated Versus Standard Treatment for Patients With a First Episode of Psychotic Illness" (*British Medical Journal*, Vol. 331, (7517):602). Part of the study included a questionnaire that was designed to measure client satisfaction for both the integrated treatment and a standard treatment. The data on the WeissStats CD are based on the results of the client questionnaire.

3.92 The Etruscans. Anthropologists are still trying to unravel the mystery of the origins of the Etruscan empire, a highly advanced Italic civilization formed around the eighth century B.C. in central Italy. Were they native to the Italian peninsula or, as many aspects of their civilization suggest, did they migrate from the East by land or sea? The maximum head breadth, in millimeters, of 70 modern Italian male skulls and that of 84 preserved Etruscan male skulls were analyzed to help researchers decide whether the Etruscans were native to Italy. The resulting data can be found on the WeissStats CD. [SOURCE: N. Barnicot and D. Brothwell, "The Evaluation of Metrical Data in the Comparison of Ancient and Modern Bones." In *Medical Biology and Etruscan Origins*, G. Wolstenholme and C. O'Connor, eds., Little, Brown & Co., 1959]

Extending the Concepts and Skills

3.93 Outliers. In Exercise 3.54 on page 101, we discussed *outliers,* or observations that fall well outside the overall pattern of the data. The following table contains two data sets. Data Set II was obtained by removing the outliers from Data Set I.

Data Set I					Data Set II			
0	12	14	15	23	10	14	15	17
0	14	15	16	24	12	14	15	
10	14	15	17		14	15	16	

a. Compute the sample standard deviation of each of the two data sets.
b. Compute the range of each of the two data sets.
c. What effect do outliers have on variation? Explain your answer.

Grouped-Data Formulas. When data are grouped in a frequency distribution, we use the following formulas to obtain the sample mean and sample standard deviation.

Grouped-Data Formulas

$$\bar{x} = \frac{\Sigma x_i f_i}{n} \quad \text{and} \quad s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2 f_i}{n-1}},$$

where x_i denotes either class mark or midpoint, f_i denotes class frequency, and n $(= \Sigma f_i)$ denotes sample size. The sample standard deviation can also be obtained by using the computing formula

$$s = \sqrt{\frac{\Sigma x_i^2 f_i - (\Sigma x_i f_i)^2/n}{n-1}}.$$

In general, these formulas yield only approximations to the actual sample mean and sample standard deviation. We ask you to apply the grouped-data formulas in Exercises 3.94 and 3.95.

3.94 Weekly Salaries. In the following table, we repeat the salary data in Data Set II from Example 3.1.

300	300	940	450	400
400	300	300	1050	300

a. Use Definitions 3.4 and 3.6 on pages 95 and 105, respectively, to obtain the sample mean and sample standard deviation of this (ungrouped) data set.

b. A frequency distribution for Data Set II, using single-value grouping, is presented in the first two columns of the following table. The third column of the table is for the xf-values, that is, class mark or midpoint (which here is the same as the class) times class frequency. Complete the missing entries in the table and then use the grouped-data formula to obtain the sample mean.

Salary x	Frequency f	Salary · Frequency xf
300	5	1500
400	2	
450	1	
940	1	
1050	1	

c. Compare the answers that you obtained for the sample mean in parts (a) and (b). Explain why the grouped-data formula always yields the actual sample mean when the data are grouped by using single-value grouping. (*Hint:* What does xf represent for each class?)
d. Construct a table similar to the one in part (b) but with columns for x, f, $x - \bar{x}$, $(x - \bar{x})^2$, and $(x - \bar{x})^2 f$. Use the table and the grouped-data formula to obtain the sample standard deviation.
e. Compare your answers for the sample standard deviation in parts (a) and (d). Explain why the grouped-data formula always yields the actual sample standard deviation when the data are grouped by using single-value grouping.

3.95 Days to Maturity. The first two columns of the following table provide a frequency distribution, using limit grouping, for the days to maturity of 40 short-term investments, as found in *BARRON'S*. The third column shows the class marks.

Days to maturity	Frequency f	Class mark x
30–39	3	34.5
40–49	1	44.5
50–59	8	54.5
60–69	10	64.5
70–79	7	74.5
80–89	7	84.5
90–99	4	94.5

a. Use the grouped-data formulas to estimate the sample mean and sample standard deviation of the days-to-maturity data. Round your final answers to one decimal place.
b. The following table gives the raw days-to-maturity data.

70	64	99	55	64	89	87	65
62	38	67	70	60	69	78	39
75	56	71	51	99	68	95	86
57	53	47	50	55	81	80	98
51	36	63	66	85	79	83	70

Using Definitions 3.4 and 3.6 on pages 95 and 105, respectively, gives the true sample mean and sample standard deviation of the days-to-maturity data as 68.3 and 16.7, respectively, rounded to one decimal place. Compare these actual values of $\bar{x}$ and s to the estimates from part (a). Explain why the grouped-data formulas generally yield only approximations to the sample mean and sample standard deviation for non–single-value grouping.

Chebychev's Rule. A more precise version of the three-standard-deviations rule (Key Fact 3.2 on page 108) can be obtained from **Chebychev's rule,** which we state as follows.

Chebychev's Rule

For any data set and any real number $k > 1$, at least $100(1 - 1/k^2)\%$ of the observations lie within k standard deviations to either side of the mean.

Two special cases of Chebychev's rule are applied frequently, namely, when $k = 2$ and $k = 3$. These state, respectively, that:

- At least 75% of the observations in any data set lie within two standard deviations to either side of the mean.
- At least 89% of the observations in any data set lie within three standard deviations to either side of the mean.

Exercises 3.96–3.99 concern Chebychev's rule.

3.96 Verify that the two statements in the preceding bulleted list are indeed special cases of Chebychev's rule with $k = 2$ and $k = 3$, respectively.

3.97 Consider the data sets portrayed in Figs. 3.5 and 3.6 on page 108.

a. Chebychev's rule says that at least 75% of the observations lie within two standard deviations to either side of the mean. What percentage of the observations portrayed in Fig. 3.5 actually lie within two standard deviations to either side of the mean?

b. Chebychev's rule says that at least 89% of the observations lie within three standard deviations to either side of the mean. What percentage of the observations portrayed in Fig. 3.5 actually lie within three standard deviations to either side of the mean?

c. Repeat parts (a) and (b) for the data portrayed in Fig. 3.6.

d. From parts (a)–(c), we see that Chebychev's rule provides a lower bound on, rather than a precise estimate for, the percentage of observations that lie within a specified number of standard deviations to either side of the mean. Nonetheless, Chebychev's rule is quite important for several reasons. Can you think of some?

3.98 Exam Scores. Consider the following sample of exam scores, arranged in increasing order.

28	57	58	64	69	74
79	80	83	85	85	87
87	89	89	90	92	93
94	94	95	96	96	97
97	97	97	98	100	100

Note: The sample mean and sample standard deviation of these exam scores are, respectively, 85 and 16.1.

a. Use Chebychev's rule to obtain a lower bound on the percentage of observations that lie within two standard deviations to either side of the mean.

b. Use the data to obtain the exact percentage of observations that lie within two standard deviations to either side of the mean. Compare your answer here to that in part (a).

c. Use Chebychev's rule to obtain a lower bound on the percentage of observations that lie within three standard deviations to either side of the mean.

d. Use the data to obtain the exact percentage of observations that lie within three standard deviations to either side of the mean. Compare your answer here to that in part (c).

3.99 Book Costs. Chebychev's rule also permits you to make pertinent statements about a data set when only its mean and standard deviation are known. Here is an example of that use of Chebychev's rule. Information Today, Inc. publishes information on costs of new books in *The Bowker Annual Library and Book Trade Almanac*. A sample of 40 sociology books has a mean cost of \$106.75 and a standard deviation of \$10.42. Use this information and the two aforementioned special cases of Chebychev's rule to complete the following statements.

a. At least 30 of the 40 sociology books cost between ______ and ______.

b. At least ______ of the 40 sociology books cost between \$75.49 and \$138.01.

The Empirical Rule. For data sets with approximately bell-shaped distributions, we can improve on the estimates given by Chebychev's rule by using the **empirical rule,** which is as follows.

Empirical Rule

For any data set having roughly a bell-shaped distribution:

- Approximately 68% of the observations lie within one standard deviation to either side of the mean.
- Approximately 95% of the observations lie within two standard deviations to either side of the mean.
- Approximately 99.7% of the observations lie within three standard deviations to either side of the mean.

Exercises 3.100–3.103 concern the empirical rule.

3.100 In this exercise, you will compare Chebychev's rule and the empirical rule.

a. Compare the estimates given by the two rules for the percentage of observations that lie within two standard deviations to either side of the mean. Comment on the differences.

b. Compare the estimates given by the two rules for the percentage of observations that lie within three standard deviations to either side of the mean. Comment on the differences.

3.101 Malnutrition and Poverty. R. Reifen et al. studied various nutritional measures of Ethiopian school children and published their findings in the paper "Ethiopian-Born and Native Israeli School Children Have Different Growth Patterns" (*Nutrition*, Vol. 19, pp. 427–431). The study, conducted in Azezo, North West Ethiopia, found that malnutrition is prevalent in primary and secondary school children because of economic poverty. The weights, in kilograms (kg), of 60 randomly selected male Ethiopian-born school children ages 12–15 years old are presented in increasing order in the following table.

36.3	37.7	38.0	38.8	38.9	39.0	39.3	40.9	41.1
41.3	41.5	41.8	42.0	42.0	42.1	42.5	42.5	42.8
42.9	43.3	43.4	43.5	44.0	44.4	44.7	44.8	45.2
45.2	45.2	45.4	45.5	45.7	45.9	45.9	46.2	46.3
46.5	46.6	46.8	47.2	47.4	47.5	47.8	47.9	48.1
48.2	48.3	48.4	48.5	48.6	48.9	49.1	49.2	49.5
50.9	51.4	51.8	52.8	53.8	56.6			

Note: The sample mean and sample standard deviation of these weights are, respectively, 45.30 kg and 4.16 kg.

a. Use the empirical rule to estimate the percentages of the observations that lie within one, two, and three standard deviations to either side of the mean.

b. Use the data to obtain the exact percentages of the observations that lie within one, two, and three standard deviations to either side of the mean.

c. Compare your answers in parts (a) and (b).

d. A histogram for these weights is shown in Exercise 2.100 on page 76. Based on that histogram, comment on your comparisons in part (c).

e. Is it appropriate to use the empirical rule for these data? Explain your answer.

3.102 Exam Scores. Refer to the exam scores displayed in Exercise 3.98.

a. Use the empirical rule to estimate the percentages of the observations that lie within one, two, and three standard deviations to either side of the mean.

b. Use the data to obtain the exact percentages of the observations that lie within one, two, and three standard deviations to either side of the mean.

c. Compare your answers in parts (a) and (b).

d. Construct a histogram or a stem-and-leaf diagram for the exam scores. Based on your graph, comment on your comparisons in part (c).

e. Is it appropriate to use the empirical rule for these data? Explain your answer.

3.103 Book Costs. Refer to Exercise 3.99. Assuming that the distribution of costs for the 40 sociology books is approximately bell shaped, apply the empirical rule to complete the following statements, and compare your answers to those obtained in Exercise 3.99, where Chebychev's rule was used.

a. Approximately 38 of the 40 sociology books cost between ______ and ______.

b. Approximately ______ of the 40 sociology books cost between $75.49 and $138.01.

3.3 The Five-Number Summary; Boxplots

So far, we have focused on the mean and standard deviation to measure center and variation. We now examine several descriptive measures based on percentiles.

Unlike the mean and standard deviation, descriptive measures based on percentiles are *resistant*—they are not sensitive to the influence of a few extreme observations. For this reason, descriptive measures based on percentiles are often preferred over those based on the mean and standard deviation.

Quartiles

As you learned in Section 3.1, the median of a data set divides the data into two equal parts: the bottom 50% and the top 50%. The **percentiles** of a data set divide it into hundredths, or 100 equal parts. A data set has 99 percentiles, denoted $P_1, P_2, \ldots, P_{99}$. Roughly speaking, the first percentile, P_1, is the number that divides the bottom 1% of the data from the top 99%; the second percentile, P_2, is the number that divides the bottom 2% of the data from the top 98%; and so on. Note that the median is also the 50th percentile.

Certain percentiles are particularly important: the **deciles** divide a data set into tenths (10 equal parts), the **quintiles** divide a data set into fifths (5 equal parts), and the **quartiles** divide a data set into quarters (4 equal parts).

Quartiles are the most commonly used percentiles. A data set has three quartiles, which we denote Q_1, Q_2, and Q_3. Roughly speaking, the **first quartile, Q_1,** is the number that divides the bottom 25% of the data from the top 75%; the **second quartile, Q_2,** is the median, which, as you know, is the number that divides the bottom 50% of the data from the top 50%; and the **third quartile, Q_3,** is the number that divides the bottom 75% of the data from the top 25%. Note that the first and third quartiles are the 25th and 75th percentiles, respectively.

Figure 3.7 depicts the quartiles for uniform, bell-shaped, right-skewed, and left-skewed distributions.