CHAPTER 3 Numerically Summarizing Data

<table>
<thead>
<tr>
<th>Age</th>
<th>Male Resident Pop</th>
<th>Female Resident Pop</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–9</td>
<td>20,700,000</td>
<td>19,826,000</td>
</tr>
<tr>
<td>10–19</td>
<td>21,369,000</td>
<td>20,475,000</td>
</tr>
<tr>
<td>20–29</td>
<td>21,417,000</td>
<td>21,355,000</td>
</tr>
<tr>
<td>30–39</td>
<td>19,455,000</td>
<td>20,011,000</td>
</tr>
<tr>
<td>40–49</td>
<td>20,839,000</td>
<td>21,532,000</td>
</tr>
<tr>
<td>50–59</td>
<td>20,785,000</td>
<td>22,058,000</td>
</tr>
<tr>
<td>60–69</td>
<td>14,739,000</td>
<td>16,362,000</td>
</tr>
<tr>
<td>70–79</td>
<td>7,641,000</td>
<td>9,474,000</td>
</tr>
<tr>
<td>≥ 80</td>
<td>4,230,000</td>
<td>6,561,000</td>
</tr>
</tbody>
</table>

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?

14. Age of Mother The following data represent the age of the mother at childbirth for 1980 and 2013.

<table>
<thead>
<tr>
<th>Age</th>
<th>1980 Births (thousands)</th>
<th>2013 Births (thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10–14</td>
<td>9.8</td>
<td>3.1</td>
</tr>
<tr>
<td>15–19</td>
<td>551.9</td>
<td>274.6</td>
</tr>
<tr>
<td>20–24</td>
<td>1226.4</td>
<td>902.1</td>
</tr>
<tr>
<td>25–29</td>
<td>1108.2</td>
<td>1127.6</td>
</tr>
<tr>
<td>30–34</td>
<td>549.9</td>
<td>1044.0</td>
</tr>
<tr>
<td>35–39</td>
<td>140.7</td>
<td>487.5</td>
</tr>
<tr>
<td>40–44</td>
<td>23.2</td>
<td>110.3</td>
</tr>
<tr>
<td>45–49</td>
<td>1.1</td>
<td>8.2</td>
</tr>
</tbody>
</table>

Source: National Vital Statistics Reports

(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 2013.
(c) Which year has the higher mean age?
(d) Which year has more dispersion in age?

15. Approximate the median of the frequency distribution in Problem 2.
16. Approximate the median of the frequency distribution in Problem 4.

Problems 15 and 16 use the following steps to approximate the median from grouped data.

### Approximating the Median from Grouped Data

**Step 1** Construct a cumulative frequency distribution.

**Step 2** Identify the class in which the median lies. Remember, the median can be obtained by determining the observation that lies in the middle.

**Step 3** Interpolate the median using the formula

\[
\text{Median} = M = L + \frac{n - CF}{f} \times i
\]

where
- \(L\) is the lower class limit of the class containing the median
- \(n\) is the number of data values in the frequency distribution
- \(CF\) is the cumulative frequency of the class immediately preceding the class containing the median
- \(f\) is the frequency of the median class
- \(i\) is the class width of the class containing the median

17. Determine the modal class of the frequency distribution in Problem 1.
18. Determine the modal class of the frequency distribution in Problem 2.

### 3.4 Measures of Position and Outliers

**Objectives**

1. Determine and interpret z-scores
2. Interpret percentiles
3. Determine and interpret quartiles
4. Determine and interpret the interquartile range
5. Check a set of data for outliers

In Section 3.1, we determined measures of central tendency, which 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, which describe the relative position of a certain data value within the entire set of data.
Determine and Interpret z-Scores

At the end of the 2014 season, the Los Angeles Angels led the American League with 773 runs scored, while the Colorado Rockies led the National League with 755 runs scored. It appears that the Angels are the better run-producing team. However, this comparison is unfair because the two teams play in different leagues. The Angels play in the American League, where the designated hitter bats for the pitcher, whereas the Rockies 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. We can do this 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. We find it 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:

\[
\begin{align*}
\text{Population z-Score} & \quad z = \frac{x - \mu}{\sigma} \\
\text{Sample z-Score} & \quad z = \frac{x - \bar{x}}{s}
\end{align*}
\]

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

**In Other Words**
The z-score provides a way to compare apples to oranges by converting variables with different centers or spreads to variables with the same center (0) and spread (1).

**EXAMPLE 1** Comparing z-Scores

**Problem** Determine whether the Los Angeles Angels or the Colorado Rockies had a relatively better run-producing season. The Angels scored 773 runs and play in the American League, where the mean number of runs scored was \( \mu = 677.4 \) and the standard deviation was \( \sigma = 51.7 \) runs. The Rockies scored 755 runs and play in the National League, where the mean number of runs scored was \( \mu = 640.0 \) and the standard deviation was \( \sigma = 55.9 \) runs.

**Approach** To determine which team had the relatively better run-producing season, 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, compute the population z-score.

**Solution** We compute each team’s z-score, rounded to two decimal places.

Angels: 
\[
z \text{-score} = \frac{x - \mu}{\sigma} = \frac{773 - 677.4}{51.7} = 1.85
\]

Rockies: 
\[
z \text{-score} = \frac{x - \mu}{\sigma} = \frac{755 - 640.0}{55.9} = 2.06
\]

So the Angels had run production 1.85 standard deviations above the mean, while the Rockies had run production 2.06 standard deviations above the mean. Therefore, the Rockies had a relatively better year at scoring runs than the Angels.
In Example 1, the team with the higher $z$-score was said to have a relatively better season in producing runs. With negative $z$-scores, we need to be careful when deciding the better outcome. For example, suppose Bob and Mary run a marathon. If Bob finished the marathon in 213 minutes, where the mean finishing time among all men was 242 minutes with a standard deviation of 57 minutes, and Mary finished the marathon in 241 minutes, where the mean finishing time among all women was 273 minutes with a standard deviation of 52 minutes, who did better in the race? Since Bob’s $z$-score is $z_{Bob} = \frac{213 - 242}{57} = -0.51$ and Mary’s $z$-score is $z_{Mary} = \frac{241 - 273}{52} = -0.62$, Mary did better. Even though Bob’s $z$-score is larger, Mary did better because she is more standard deviations below the mean.

2 Interpret Percentiles

Recall that the median divides the lower 50% of a set of data from the upper 50%. The median is a special case of a general concept called the percentiles.

Definition

The $k$th percentile, denoted $P_k$, of a set of data is a value such that $k$ percent of the observations are less than or equal to the value.

So percentiles divide a set of data that is written in ascending order into 100 parts; thus 99 percentiles can be determined. For example, $P_1$ divides the bottom 1% of the observations from the top 99%, $P_2$ divides the bottom 2% of the observations from the top 98%, and so on. Figure 17 displays the 99 possible percentiles.

Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT college entrance exam, use percentiles to let students know how they scored on the exam in relation to all other students who took the exam.

EXAMPLE 2 Interpret a Percentile

Problem Jennifer just received the results of her SAT exam. Her SAT Mathematics score of 600 is in the 74th percentile. What does this mean?

Approach The $k$th percentile of an observation means that $k$ percent of the observations are less than or equal to the observation.

Interpretation A percentile rank of 74% means that 74% of SAT Mathematics scores are less than or equal to 600 and 26% of the scores are greater. So 26% of the students who took the exam scored better than Jennifer.

3 Determine and Interpret Quartiles

The most common percentiles are quartiles. Quartiles divide data sets into fourths, or four equal parts.
SECTION 3.4 Measures of Position and Outliers

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, $P_{75}$.

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.

Finding Quartiles

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

**Step 2** Determine the median, $M$, or second quartile, $Q_2$.

**Step 3** Divide the data set into halves: the observations below (to the left of) $M$ and the observations above $M$. The first quartile, $Q_1$, is the median of the bottom half of the data and the third quartile, $Q_3$, is the median of the top half of the data.

EXAMPLE 3 Finding and Interpreting Quartiles

**Problem** The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual’s vehicle. The data in Table 16 represent a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute. Find and interpret the first, second, and third quartiles for collision coverage claims.

**Table 16**

<table>
<thead>
<tr>
<th>Smallest Data Value</th>
<th>Median</th>
<th>Largest Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>25% of the data</td>
<td>25% of the data</td>
<td>25% of the data</td>
</tr>
<tr>
<td>$6751$</td>
<td>$9908$</td>
<td>$2332$</td>
</tr>
<tr>
<td>$189$</td>
<td>$1185$</td>
<td>$1953$</td>
</tr>
<tr>
<td>$10,034$</td>
<td>$735$</td>
<td>$1657$</td>
</tr>
</tbody>
</table>

**Approach** Follow the steps given above.

**Solution**

**Step 1** The data written in ascending order are given as follows:

- $180, 189, 370, 618, 735, 802, 1185, 1414, 1657, 1953, 21,147, 2332, 2336, 3461, 4668, 6751, 9908, 10,034$

**Step 2** There are $n = 18$ observations, so the median, or second quartile, $Q_2$, is the mean of the 9th and 10th observations. Therefore, $M = Q_2 = \frac{1657 + 1953}{2} = 1805$.

**Step 3** The median of the bottom half of the data is the first quartile, $Q_1$. As shown next, the median of these data is the 5th observation, so $Q_1 = 735$.

- 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, $Q_2$, divides the bottom 50% of the data from the top 50%; it is equivalent to the 50th percentile or the median.
- The third quartile, $Q_3$, divides the bottom 75% of the data from the top 25%; it is equivalent to the 75th percentile.

Figure 18 illustrates the concept of quartiles.

NOTE
If the number of observations is odd, do not include the median when determining $Q_1$ and $Q_3$ by hand.
The median of the top half of the data is the third quartile, $Q_3$. As shown next, the median of these data is the 5th observation, so $Q_3 = $4668.

\[
\begin{align*}
$1953 & \quad $2332 & \quad $2336 & \quad $3461 & \quad $4668 & \quad $6751 & \quad $9908 & \quad $10,034 & \quad $21,147 \\
\uparrow & \quad & \quad & \quad & \quad & \quad & \quad & \quad & \\
Q_3
\end{align*}
\]

**Interpretation**  Interpret the quartiles as percentiles. For example, 25% of the collision claims are less than or equal to the first quartile, $735, and 75% of the collision claims are greater than $735. Also, 50% of the collision claims are less than or equal to $1805, the second quartile, and 50% of the collision claims are greater than $1805. Finally, 75% of the collision claims are less than or equal to $4668, the third quartile, and 25% of the collision claims are greater than $4668.

**EXAMPLE 4**  Finding Quartiles Using Technology

**Problem**  Find the quartiles of the collision coverage claims data in Table 16.

**Approach**  Use both StatCrunch and Minitab to obtain the quartiles. The steps for obtaining quartiles using a TI-83/84 Plus graphing calculator, Minitab, Excel, and StatCrunch are given in the Technology Step-by-Step on pages 160–161.

**Solution**  The results obtained from StatCrunch [Figure 19(a)] agree with our “by hand” solution. In Figure 19(b), notice that the first quartile, 706, and the third quartile, 5189, reported by Minitab disagree with our “by hand” and StatCrunch result. This difference is due to the fact that StatCrunch and Minitab use different algorithms for obtaining quartiles.

**Determine and Interpret the Interquartile Range**

So far we have discussed three measures of dispersion: range, standard deviation, and variance, all of which are not resistant. Quartiles, however, are resistant. For this reason, quartiles are used to define a fourth measure of dispersion.

**Definition**  The interquartile range, IQR, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

\[\text{IQR} = Q_3 - Q_1\]

The interpretation of the interquartile range is similar to that of the range and standard deviation. That is, the more spread a set of data has, the higher the interquartile range will be.
Let’s compare the measures of central tendency and dispersion discussed thus far for the collision claim data. The mean collision claim is $3874.4 and the median is $1805. The median is more representative of the “center” because the data are skewed to the right (only 5 of the 18 observations are greater than the mean). The range is $21,147 - $180 = $20,967. The standard deviation is $5301.6 and the interquartile range is $3933. The values of the range and standard deviation are affected by the extreme claim of $21,147. In fact, if this claim had been $120,000 (let’s say the claim was for a totaled Mercedes S-class AMG), then the range and standard deviation would increase to $119,820 and $27,782.5, respectively. The interquartile range would not be affected. Therefore, when the distribution of data is highly skewed or contains extreme observations, it is best to use the interquartile range as the measure of dispersion because it is resistant.

### Example 5

**Problem** Determine and interpret the interquartile range of the collision claim data from Example 3.

**Approach** Use the quartiles found by hand in Example 3. The interquartile range, IQR, is found by computing the difference between the third and first quartiles. It represents the range of the middle 50% of the observations.

**Solution** The interquartile range is

\[
\text{IQR} = Q_3 - Q_1 = $4668 - $735 = $3933
\]

**Interpretation** The IQR, that is, the range of the middle 50% of the observations, in the collision claim data is $3933.

### Summary: Which Measures to Report

<table>
<thead>
<tr>
<th>Shape of Distribution</th>
<th>Measure of Central Tendency</th>
<th>Measure of Dispersion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetric</td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>Skewed left or skewed right</td>
<td>Median</td>
<td>Interquartile range</td>
</tr>
</tbody>
</table>

For the remainder of this text, the direction **describe the distribution** will mean to describe its shape (skewed left, skewed right, symmetric), its center (mean or median), and its spread (standard deviation or interquartile range).

---

**Check a Set of Data for Outliers**

When performing any type of data analysis, we should always check for extreme observations in the data set. Extreme observations are referred to as **outliers**. Outliers 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.

Outliers do not always occur because of error. 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 F430 Spider...
(approximately $175,000), 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.

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.

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

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

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

**EXAMPLE 6  Checking for Outliers**

**Problem** Check the collision coverage claims data in Table 16 for outliers.

**Approach** 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 found in Example 3 are \( Q_1 = \$735 \) and \( Q_3 = \$4668 \).

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

\[
IQR = Q_3 - Q_1 \\
= \$4668 - \$735 \\
= \$3933
\]

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

\[
LF = Q_1 - 1.5(IQR) \\
= \$735 - 1.5(\$3933) \\
= -\$5164.5
\]

The upper fence, UF, is

\[
UF = Q_3 + 1.5(IQR) \\
= \$4668 + 1.5(\$3933) \\
= \$10,567.5
\]

**Step 4** There are no observations below the lower fence. However, there is an observation above the upper fence. The claim of $21,147 is an outlier.

**Technology Step-by-Step  Determining Quartiles**

<table>
<thead>
<tr>
<th><strong>TI-83/84 Plus</strong></th>
<th><strong>Minitab</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Follow the same steps given to compute the mean and median from raw data. (Section 3.1)</td>
<td>Follow the same steps given to compute the mean and median from raw data. (Section 3.1)</td>
</tr>
</tbody>
</table>
75-inch man or a 70-inch woman? Standard deviation of 3.8 inches. Who is relatively taller, a 69.6 inches tall, with a standard deviation of 3.0 inches, while a 67-inch man or a 62-inch woman?

6. Birth Weights Babies born after a gestation period of 32–35 weeks have a mean weight of 2600 grams and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 2400 grams and a 40-week gestation period baby weighs 3300 grams. What is the z-score for the 34-week gestation period baby? What is the z-score for the 40-week gestation period baby?

Which baby weighs less relative to the gestation period?

Vocabulary

1. The _______ represents the number of standard deviations an observation is from the mean.
2. The _______ _______ of a data set is a value such that k percent of the observations are less than or equal to the value.
3. _______ divide data sets into fourths.
4. The _______ _______ is the range of the middle 50% of the observations in a data set.

Applying the Concepts

5. Birth Weights Babies born after a gestation period of 32–35 weeks have a mean weight of 2600 grams and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 2400 grams and a 40-week gestation period baby weighs 3300 grams. What is the z-score for the 34-week gestation period baby? What is the z-score for the 40-week gestation period baby? Which baby weighs less relative to the gestation period?

6. Birth Weights Babies born after a gestation period of 32–35 weeks have a mean weight of 2600 grams and a standard deviation of 660 grams. Babies born after a gestation period of 40 weeks have a mean weight of 3500 grams and a standard deviation of 470 grams. Suppose a 34-week gestation period baby weighs 3000 grams and a 40-week gestation period baby weighs 3900 grams. What is the z-score for the 34-week gestation period baby? What is the z-score for the 40-week gestation period baby? Which baby weighs less relative to the gestation period?

7. Men versus Women The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 3.0 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 3.8 inches. Who is relatively taller, a 75-inch man or a 70-inch woman?

Source: CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005

8. Men versus Women The average 20- to 29-year-old man is 69.6 inches tall, with a standard deviation of 3.0 inches, while the average 20- to 29-year-old woman is 64.1 inches tall, with a standard deviation of 3.8 inches. Who is relatively taller, a 67-inch man or a 62-inch woman?

Source: CDC Vital and Health Statistics, Advance Data, Number 361, July 5, 2005

9. ERA Champions In 2014, Clayton Kershaw of the Los Angeles Dodgers had the lowest earned-run average (ERA is the mean number of runs yielded per nine innings pitched) of any starting pitcher in the National League, with an ERA of 1.77. Also in 2014, Felix Hernandez of the Seattle Mariners had the lowest ERA of any starting pitcher in the American League with an ERA of 2.14. In the National League, the mean ERA in 2014 was 3.430 and the standard deviation was 0.721. In the American League, the mean ERA in 2014 was 3.598 and the standard deviation was 0.762. Who had the better year relative to his peers? Why?

10. Batting Champions The highest batting average ever recorded in Major League Baseball was by Ted Williams in 1941 when he hit 0.406. That year, the mean and standard deviation for batting average were 0.2806 and 0.0328. In 2014, Jose Altuve was the American League batting champion, with a batting average of 0.341. In 2014, the mean and standard deviation for batting average were 0.2679 and 0.0282. Who had the better year relative to his peers, Williams or Altuve? Why?

11. Swim Ryan Murphy, nephew of the author, swims for the University of California at Berkeley. Ryan's best time in the 100-meter backstroke is 45.3 seconds. The mean of all NCAA swimmers in this event is 48.62 seconds with a standard deviation of 0.98 second. Ryan's best time in the 200-meter backstroke is 99.32 seconds. The mean of all NCAA swimmers in this event is 106.58 seconds with a standard deviation of 2.38 seconds. In which race is Ryan better?

12. Triathlon Roberto finishes a triathlon (750-meter swim, 5-kilometer run, and 20-kilometer bicycle) in 63.2 minutes. Among all men in the race, the mean finishing time was 69.4 minutes with a standard deviation of 8.9 minutes. Zandra finishes the same triathlon in 79.3 minutes. Among all women in the race, the mean finishing time was 84.7 minutes with a standard deviation of 7.4 minutes. Who did better in relation to their gender?

13. School Admissions A highly selective boarding school will only admit students who place at least 1.5 standard deviations above the mean on a standardized test that has a mean of 200 and a standard deviation of 26. What is the minimum score that an applicant must make on the test to be accepted?

14. Quality Control A manufacturer of bolts has a quality-control policy that requires it to destroy any bolts that are more than 2 standard deviations from the mean. The quality-control engineer knows that the bolts coming off the assembly line have...
a mean length of 8 cm with a standard deviation of 0.05 cm. For what lengths will a bolt be destroyed?

**NW 15. You Explain It! Percentiles** Explain the meaning of the following percentiles.

*Source: Advance Data from Vital and Health Statistics*

(a) The 15th percentile of the head circumference of males 3 to 5 months of age is 41.0 cm.
(b) The 90th percentile of the waist circumference of females 2 years of age is 52.7 cm.
(c) Anthropometry involves the measurement of the human body. One goal of these measurements is to assess how body measurements may be changing over time. The following table represents the standing height of males aged 20 years or older for various age groups. Based on the percentile measurements of the different age groups, what might you conclude?

<table>
<thead>
<tr>
<th>Age</th>
<th>10th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>20–29</td>
<td>166.8</td>
<td>171.5</td>
<td>176.7</td>
<td>181.4</td>
<td>186.8</td>
</tr>
<tr>
<td>30–39</td>
<td>166.9</td>
<td>171.3</td>
<td>176.0</td>
<td>181.9</td>
<td>186.2</td>
</tr>
<tr>
<td>40–49</td>
<td>167.9</td>
<td>172.1</td>
<td>176.9</td>
<td>182.1</td>
<td>186.0</td>
</tr>
<tr>
<td>50–59</td>
<td>166.0</td>
<td>170.8</td>
<td>176.0</td>
<td>181.2</td>
<td>185.4</td>
</tr>
<tr>
<td>60–69</td>
<td>165.3</td>
<td>170.1</td>
<td>175.1</td>
<td>179.5</td>
<td>183.7</td>
</tr>
<tr>
<td>70–79</td>
<td>163.2</td>
<td>167.5</td>
<td>172.9</td>
<td>178.1</td>
<td>181.7</td>
</tr>
<tr>
<td>80 or older</td>
<td>161.7</td>
<td>166.1</td>
<td>170.5</td>
<td>175.3</td>
<td>179.4</td>
</tr>
</tbody>
</table>

**16. You Explain It! Percentiles** Explain the meaning of the following percentiles.

*Source: National Center for Health Statistics.*

(a) The 5th percentile of the weight of males 36 months of age is 12.0 kg.
(b) The 95th percentile of the length of newborn females is 53.8 cm.

**17. You Explain It! Quartiles** Violent crimes include rape, robbery, assault, and homicide. The following is a summary of the violent-crime rate (violent crimes per 100,000 population) for all 50 states in the United States plus Washington, D.C., in 2012.

\[ Q_1 = 252.4 \quad Q_2 = 333.8 \quad Q_3 = 454.5 \]

(a) Provide an interpretation of these results.
(b) Determine and interpret the interquartile range.
(c) The violent-crime rate in Washington, D.C., in 2012 was 1243.7. Would this be an outlier?
(d) Do you believe that the distribution of violent-crime rates is skewed or symmetric? Why?

**18. You Explain It! Quartiles** One variable that is measured by online homework systems is the amount of time a student spends on homework for each section of the text. The following is a summary of the number of minutes a student spends for each section of the text for the fall 2014 semester in a College Algebra class at Joliet Junior College.

\[ Q_1 = 42 \quad Q_2 = 51.5 \quad Q_3 = 72.5 \]

(a) Provide an interpretation of these results.
(b) Determine and interpret the interquartile range.
(c) Suppose a student spent 2 hours doing homework for a section. Is this an outlier?

**NW 19. Ogives and Percentiles** The following graph is an ogive of IQ scores. The vertical axis in an ogive is the cumulative relative frequency and can also be interpreted as a percentile.

**Percentile Ranks of IQ Scores**

<table>
<thead>
<tr>
<th>Percentile</th>
<th>10th</th>
<th>25th</th>
<th>50th</th>
<th>75th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>60</td>
<td>80</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Find and interpret the percentile rank of an individual whose IQ is 100.
(b) Find and interpret the percentile rank of an individual whose IQ is 120.
(c) What score corresponds to the 60th percentile for IQ?

**20. Ogives and Percentiles** The following graph is an ogive of the mathematics scores on the SAT. The vertical axis in an ogive is the cumulative relative frequency and can also be interpreted as a percentile.

**SAT Mathematics Scores**

<table>
<thead>
<tr>
<th>Score</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.2</td>
</tr>
<tr>
<td>300</td>
<td>0.4</td>
</tr>
<tr>
<td>400</td>
<td>0.6</td>
</tr>
<tr>
<td>500</td>
<td>0.8</td>
</tr>
<tr>
<td>600</td>
<td>1.0</td>
</tr>
</tbody>
</table>

(a) Find and interpret the percentile rank of a student who scored 450 on the SAT mathematics exam.
(b) Find and interpret the percentile rank of a student who scored 750 on the SAT mathematics exam.
(c) If Jane scored at the 44th percentile, what was her score?

**NW 21. SMART Car** The following data represent the miles per gallon of a random sample of SMART cars with a three-cylinder, 1.0-liter engine.
(a) Compute the \( z \)-score corresponding to the individual who obtained 36.3 miles per gallon. Interpret this result.
(b) Determine the quartiles.
(c) Compute and interpret the interquartile range, IQR.
(d) Determine the lower and upper fences. Are there any outliers?

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

<table>
<thead>
<tr>
<th>5.7</th>
<th>8.9</th>
<th>9.6</th>
<th>10.6</th>
<th>11.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.7</td>
<td>9.4</td>
<td>9.9</td>
<td>10.7</td>
<td>12.9</td>
</tr>
<tr>
<td>7.8</td>
<td>9.5</td>
<td>10.0</td>
<td>11.0</td>
<td>13.0</td>
</tr>
<tr>
<td>8.7</td>
<td>9.6</td>
<td>10.3</td>
<td>11.2</td>
<td>13.4</td>
</tr>
</tbody>
</table>

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 and interpret the interquartile range, IQR.
(d) Determine the lower and upper fences. Are there any outliers?

23. Rate of Return of Google  The following data represent the monthly rate of return of Google common stock from its inception in January 2007 through November 2014.

\[
\begin{align*}
-0.10 & -0.02 & 0.00 & 0.02 & -0.10 & 0.03 & 0.04 & -0.15 & -0.08 \\
0.02 & 0.01 & -0.18 & -0.10 & -0.18 & 0.14 & 0.07 & -0.01 & 0.09 \\
0.03 & 0.10 & -0.17 & -0.10 & 0.05 & 0.05 & 0.08 & 0.08 & -0.07 \\
0.06 & 0.25 & -0.07 & -0.02 & 0.10 & 0.01 & 0.09 & -0.07 & 0.17 \\
0.05 & -0.02 & 0.30 & -0.14 & 0.00 & 0.05 & 0.06 & -0.08 & 0.17
\end{align*}
\]

Source: Yahoo!Finance

(a) Determine and interpret the quartiles.
(b) Check the data set for outliers.

24. CO\(_2\) Emissions  The following data represent the carbon dioxide emissions from the consumption of energy per capita (total carbon dioxide emissions, in tons, divided by total population) for the countries of Europe.

\[
\begin{align*}
1.31 & 5.38 & 10.36 & 5.73 & 3.57 & 5.40 & 6.24 \\
8.59 & 9.46 & 6.48 & 11.06 & 7.94 & 4.63 & 6.12 \\
14.87 & 9.94 & 10.06 & 10.71 & 15.86 & 6.93 & 3.58 \\
4.09 & 9.91 & 161.57 & 7.82 & 8.70 & 8.33 & 9.38 \\
7.31 & 16.75 & 9.95 & 23.87 & 7.76 & 8.86 \\
\end{align*}
\]

Source: Carbon Dioxide Information Analysis Center

(a) Determine and interpret the quartiles.
(b) Is the observation corresponding to Albania, 1.31, an outlier?

25. Fraud Detection  As part of its “Customers First” program, a cellular phone company monitors monthly phone usage. The program identifies unusual use and alerts the customer that their phone may have been used by another person. The data below represent the monthly phone use in minutes of a customer enrolled in this program for the past 20 months. 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?

\[
\begin{align*}
346 & 345 & 489 & 358 & 471 \\
442 & 466 & 505 & 466 & 372 \\
442 & 461 & 515 & 549 & 437 \\
480 & 490 & 429 & 470 & 516
\end{align*}
\]

26. Stolen Credit Card  A credit card company has a fraud-detection service that determines if a card has any unusual activity. The company maintains a database of daily charges on a customer’s credit card. Days when the card was inactive are 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.

\[
\begin{align*}
143 & 166 & 113 & 188 & 133 \\
90 & 89 & 98 & 95 & 112 \\
111 & 79 & 46 & 20 & 112 \\
70 & 174 & 68 & 101 & 212
\end{align*}
\]

27. Student Survey of Income  A survey of 50 randomly selected full-time Joliet Junior College students was conducted during the Fall 2015 semester. In the survey, the students were asked to disclose their weekly income from employment. If the student did not work, $0 was entered.

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

(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.

28. Student Survey of Entertainment Spending  A survey of 40 randomly selected full-time Joliet Junior College students was conducted in the Fall 2015 semester. In the survey, the students were asked to disclose their weekly spending on entertainment. The results of the survey are as follows:

\[
\begin{align*}
21 & 54 & 64 & 33 & 65 & 32 & 21 & 16 \\
22 & 39 & 67 & 54 & 22 & 51 & 26 & 14 \\
115 & 7 & 80 & 59 & 20 & 33 & 13 & 36 \\
36 & 10 & 12 & 101 & 1000 & 26 & 38 & 8 \\
28 & 28 & 75 & 50 & 27 & 35 & 9 & 48
\end{align*}
\]
(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.
29. Pulse Rate Use the results of Problem 21 in Section 3.1 and Problem 19 in Section 3.2 to compute the z-scores for all the students. Compute the mean and standard deviation of these z-scores.
30. Travel Time Use the results of Problem 22 in Section 3.1 and Problem 20 in Section 3.2 to compute the z-scores for all the students. Compute the mean and standard deviation of these z-scores.
31. Fraud Detection Revisited Use the fraud-detection data from Problem 25 to do the following.
(a) Determine the standard deviation and interquartile range of the data.
(b) Suppose the month in which the customer used 346 minutes was not actually that customer’s phone. That particular month, the customer did not use her phone at all, so 0 minutes were used. How does changing the observation from 346 to 0 affect the standard deviation and interquartile range? What property does this illustrate?

### Explaining the Concepts

32. Write a paragraph that explains the meaning of percentiles.
33. 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?
34. 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 10% of its investment class; a five-star mutual fund is at the 90th percentile of its investment class. Interpret the meaning of a five-star mutual fund.
35. When outliers are discovered, should they always be removed from the data set before further analysis?
36. 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.
37. Explain the advantage of using z-scores to compare observations from two different data sets.
38. Explain the circumstances for which the interquartile range is the preferred measure of dispersion. What is an advantage that the standard deviation has over the interquartile range?
39. Explain what each quartile represents.

## 3.5 The Five-Number Summary and Boxplots

### Objectives

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

### 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 voted “the couple most likely to give birth to a genius.” Tukey earned his undergraduate and master's degrees in chemistry from Brown University. In 1939, he earned his doctorate in mathematics from Princeton University. 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. In December 1976, he published *Exploratory Data Analysis*, from which the following quote appears: “Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone—as the first step” (p. 3). Tukey also recommended that the 1990 Census be adjusted by means of statistical formulas. John Tukey died in New Brunswick, New Jersey, on July 26, 2000.

Let’s consider what we have learned so far. In Chapter 2, we discussed techniques for graphically representing data. These summaries included bar graphs, pie charts, histograms, stem-and-leaf plots, and time series graphs. In Sections 3.1 to 3.4, we presented techniques for measuring the center of a distribution, spread in a distribution, and relative position of observations in a distribution of data. Why do we want these summaries? What purpose do they serve?

Well, we want these summaries to see what the data can tell us. We explore the data to see if they contain interesting information that may be useful in our research. The summaries make this exploration much easier. In fact, because these summaries represent an exploration, a famous statistician named John Tukey called this material *exploratory data analysis*.

Tukey defined exploratory data analysis as “detective work—numerical detective work—or graphical detective work.” He believed exploration of data is best carried out the way a detective searches for evidence when investigating a crime. Our goal is only to collect and present evidence. Drawing conclusions (or inference) is like the deliberations of the jury. What we have done so far falls under the category of exploratory data analysis. We have only collected information and presented summaries, not reached any conclusions.

We have already seen one of Tukey’s graphical summaries, the stem-and-leaf plot. In this section, we look at two more summaries: the five-number summary and the boxplot.

### Compute the Five-Number Summary

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