CHAPTER 3

Descriptive Measures

CHAPTER OBJECTIVES

In Chapter 2, you began your study of descriptive statistics. There you learned how to organize data into tables and summarize data with graphs.

Another method of summarizing data is to compute numbers, such as averages and percentiles, that describe the data set. Numbers that are used to describe data sets are called **descriptive measures.** In this chapter, we continue our discussion of descriptive statistics by examining some of the most commonly used descriptive measures.

In Section 3.1, we present *measures of center*—descriptive measures that indicate the center, or most typical value, in a data set. Next, in Section 3.2, we examine *measures of variation*—descriptive measures that indicate the amount of variation or spread in a data set.

The five-number summary, which we discuss in Section 3.3, includes descriptive measures that can be used to obtain both measures of center and measures of variation. That summary also provides the basis for a widely used graphical display, the boxplot.

In Section 3.4, we examine descriptive measures of populations. We also illustrate how sample data can be used to provide estimates of descriptive measures of populations when census data are unavailable.

CASE STUDY

U.S. Presidential Election

From the document *Official Presidential General Election Results* published by the Federal Election Commission, we found final results of the 2008 U.S. presidential election. Barack Obama received 365 electoral votes versus 173 electoral votes obtained by John McCain. Thus, the Obama and McCain electoral vote percentages were 67.8% and 32.2%, respectively.

From a popular vote perspective, the election was much closer: Obama got 69,456,897 votes and McCain received 59,934,814 votes. Taking into account that the total popular vote for all candidates was 131,257,328, we see that the Obama and McCain popular vote percentages were 52.9% and 45.7%, respectively.

CHAPTER OUTLINE

We can gain further insight into the election results by investigating the state-by-state percentages. The following table gives us that information for Obama.

In this chapter, we demonstrate several additional techniques to help you analyze data. At the end of the chapter, you will apply those techniques to analyze the state-by-state percentages presented in the table.

State	% Obama	State	% Obama	State	% Obama
Alabama	38.7	Kentucky	41.2	North Dakota	44.6
Alaska	37.9	Louisiana	39.9	Ohio	51.5
Arizona	45.1	Maine	57.7	Oklahoma	34.4
Arkansas	38.9	Maryland	61.9	Oregon	56.7
California	61.0	Massachusetts	61.8	Pennsylvania	54.5
Colorado	53.7	Michigan	57.4	Rhode Island	62.9
Connecticut	60.6	Minnesota	54.1	South Carolina	44.9
Delaware	61.9	Mississippi	43.0	South Dakota	44.7
DC	92.5	Missouri	49.3	Tennessee	41.8
Florida	51.0	Montana	47.2	Texas	43.7
Georgia	47.0	Nebraska	41.6	Utah	34.4
Hawaii	71.8	Nevada	55.1	Vermont	67.5
Idaho	36.1	New Hampshire	54.1	Virginia	52.6
Illinois	61.9	New Jersey	57.3	Washington	57.7
Indiana	49.9	New Mexico	56.9	West Virginia	42.6
Iowa	53.9	New York	62.8	Wisconsin	56.2
Kansas	41.7	North Carolina	49.7	Wyoming	32.5

3.1 Measures of Center

Descriptive measures that indicate where the center or most typical value of a data set lies are called **measures of central tendency** or, more simply, **measures of center.** Measures of center are often called *averages.*

In this section, we discuss the three most important measures of center: the *mean, median,* and *mode.* The mean and median apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative (categorical) data.

The Mean

The most commonly used measure of center is the *mean.* When people speak of taking an average, they are most often referring to the mean.

DEFINITION 3.1

Mean of a Data Set

The **mean** of a data set is the sum of the observations divided by the number of observations.

What Does It Mean?

The mean of a data set is its arithmetic average.

EXAMPLE 3.1 The Mean

Weekly Salaries Professor Hassett spent one summer working for a small mathematical consulting firm. The firm employed a few senior consultants, who made

between \$800 and \$1050 per week; a few junior consultants, who made between \$400 and \$450 per week; and several clerical workers, who made \$300 per week.

The firm required more employees during the first half of the summer than the second half. Tables 3.1 and 3.2 list typical weekly earnings for the two halves of the summer. Find the mean of each of the two data sets.

TABLE 3.1
Data Set I

\$300	300	300	940	300
300	400	300	400	
450	800	450	1050	

TABLE 3.2
Data Set II

\$300	300	940	450	400
400	300	300	1050	300

SC Report 3.1

You try it!

Exercise 3.15(a) on page 97

Solution As we see from Table 3.1, Data Set I has 13 observations. The sum of those observations is \$6290, so

$$\text{Mean of Data Set I} = \frac{\$6290}{13} = \$483.85 \text{ (rounded to the nearest cent).}$$

Similarly,

$$\text{Mean of Data Set II} = \frac{\$4740}{10} = \$474.00.$$

Interpretation The employees who worked in the first half of the summer earned more, on average (a mean salary of \$483.85), than those who worked in the second half (a mean salary of \$474.00).

The Median

Another frequently used measure of center is the median. Essentially, the *median* of a data set is the number that divides the bottom 50% of the data from the top 50%. A more precise definition of the median follows.

DEFINITION 3.2

Median of a Data Set

Arrange the data in increasing order.

- If the number of observations is odd, then the **median** is the observation exactly in the middle of the ordered list.
- If the number of observations is even, then the **median** is the mean of the two middle observations in the ordered list.

In both cases, if we let n denote the number of observations, then the median is at position $(n + 1)/2$ in the ordered list.

What Does It Mean?

The median of a data set is the middle value in its ordered list.

EXAMPLE 3.2 The Median

Weekly Salaries Consider again the two sets of salary data shown in Tables 3.1 and 3.2. Determine the median of each of the two data sets.

Solution To find the median of Data Set I, we first arrange the data in increasing order:

300 300 300 300 300 300 **400** 400 450 450 800 940 1050

The number of observations is 13, so $(n + 1)/2 = (13 + 1)/2 = 7$. Consequently, the median is the seventh observation in the ordered list, which is 400 (shown in boldface).

To find the median of Data Set II, we first arrange the data in increasing order:

300 300 300 300 **300** **400** 400 450 940 1050

The number of observations is 10, so $(n + 1)/2 = (10 + 1)/2 = 5.5$. Consequently, the median is halfway between the fifth and sixth observations (shown in boldface) in the ordered list, which is 350.

Interpretation Again, the analysis shows that the employees who worked in the first half of the summer tended to earn more (a median salary of \$400) than those who worked in the second half (a median salary of \$350).

SC Report 3.2

You try it!

Exercise 3.15(b) on page 97

To determine the median of a data set, you must first arrange the data in increasing order. Constructing a stem-and-leaf diagram as a preliminary step to ordering the data is often helpful.

The Mode

The final measure of center that we discuss here is the *mode*.

DEFINITION 3.3

Mode of a Data Set

Find the frequency of each value in the data set.

- If no value occurs more than once, then the data set has *no mode*.
- Otherwise, any value that occurs with the greatest frequency is a **mode** of the data set.

What Does It Mean?

The mode of a data set is its most frequently occurring value.

EXAMPLE 3.3 The Mode

Weekly Salaries Determine the mode(s) of each of the two sets of salary data given in Tables 3.1 and 3.2 on page 91.

Solution Referring to Table 3.1, we obtain the frequency of each value in Data Set I, as shown in Table 3.3. From Table 3.3, we see that the greatest frequency is 6, and that 300 is the only value that occurs with that frequency. So the mode is \$300.

Proceeding in the same way, we find that, for Data Set II, the greatest frequency is 5 and that 300 is the only value that occurs with that frequency. So the mode is \$300.

TABLE 3.3
Frequency distribution for Data Set I

Salary	Frequency
300	6
400	2
450	2
800	1
940	1
1050	1

Interpretation The most frequent salary was \$300 both for the employees who worked in the first half of the summer and those who worked in the second half.

SC Report 3.3

You try it!

Exercise 3.15(c) on page 97

A data set will have more than one mode if more than one of its values occurs with the greatest frequency. For instance, suppose the first two \$300-per-week employees who worked in the first half of the summer were promoted to \$400-per-week jobs. Then the weekly earnings for the 13 employees would be as follows.

\$400	400	300	940	300
300	400	300	400	
450	800	450	1050	

Now, both the value 300 and the value 400 would occur with greatest frequency, 4. This new data set would thus have two modes, 300 and 400.

Comparison of the Mean, Median, and Mode

The mean, median, and mode of a data set are often different. Table 3.4 summarizes the definitions of these three measures of center and gives their values for Data Set I and Data Set II, which we computed in Examples 3.1–3.3.

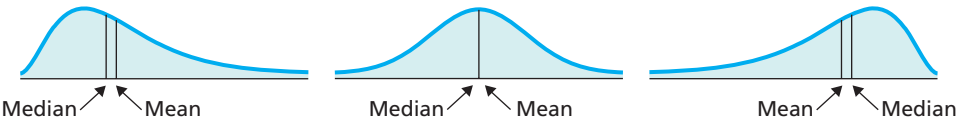
TABLE 3.4
Means, medians, and modes of salaries in Data Set I and Data Set II

Measure of center	Definition	Data Set I	Data Set II
Mean	$\frac{\text{Sum of observations}}{\text{Number of observations}}$	$483.85	$474.00
Median	Middle value in ordered list	$400.00	$350.00
Mode	Most frequent value	$300.00	$300.00

In both Data Sets I and II, the mean is larger than the median. The reason is that the mean is strongly affected by the few large salaries in each data set. In general, the mean is sensitive to extreme (very large or very small) observations, whereas the median is not. Consequently, when the choice for the measure of center is between the mean and the median, the median is usually preferred for data sets that have extreme observations.

Figure 3.1 shows the relative positions of the mean and median for right-skewed, symmetric, and left-skewed distributions. Note that the mean is pulled in the direction of skewness, that is, in the direction of the extreme observations. For a right-skewed distribution, the mean is greater than the median; for a symmetric distribution, the mean and the median are equal; and for a left-skewed distribution, the mean is less than the median.

FIGURE 3.1
Relative positions of the mean and median for (a) right-skewed, (b) symmetric, and (c) left-skewed distributions



(a) Right skewed (b) Symmetric (c) Left skewed

APPLET
Applet 3.1

A **resistant measure** is not sensitive to the influence of a few extreme observations. The median is a resistant measure of center, but the mean is not. A *trimmed mean* can improve the resistance of the mean: removing a percentage of the smallest and largest observations before computing the mean gives a **trimmed mean.** In Exercise 3.54, we discuss trimmed means in more detail.

The mode for each of Data Sets I and II differs from both the mean and the median. Whereas the mean and the median are aimed at finding the center of a data set, the mode is really not—the value that occurs most frequently may not be near the center.

It should now be clear that the mean, median, and mode generally provide different information. There is no simple rule for deciding which measure of center to use in a given situation. Even experts may disagree about the most suitable measure of center for a particular data set.

EXAMPLE 3.4 Selecting an Appropriate Measure of Center

a. A student takes four exams in a biology class. His grades are 88, 75, 95, and 100. Which measure of center is the student likely to report?

b. The National Association of REALTORS publishes data on resale prices of U.S. homes. Which measure of center is most appropriate for such resale prices?

c. The 2009 Boston Marathon had two categories of official finishers: male and female, of which there were 13,547 and 9,302, respectively. Which measure of center should be used here?

Solution

a. Chances are that the student would report the mean of his scores, which is 89.5. The mean is probably the most suitable measure of center for the student to use because it takes into account the numerical value of each score and therefore indicates his overall performance.

b. The most appropriate measure of center for resale home prices is the median because it is aimed at finding the center of the data on resale home prices and because it is not strongly affected by the relatively few homes with extremely high resale prices. Thus the median provides a better indication of the "typical" resale price than either the mean or the mode.

Exercise 3.23
on page 98

c. The only suitable measure of center for these data is the mode, which is "male." Each observation in this data set is either "male" or "female." There is no way to compute a mean or median for such data. *Of the mean, median, and mode, the mode is the only measure of center that can be used for qualitative data.*

Many measures of center that appear in newspapers or that are reported by government agencies are medians, as is the case for household income and number of years of school completed. In an attempt to provide a clearer picture, some reports include both the mean and the median. For instance, the National Center for Health Statistics does so for daily intake of nutrients in the publication *Vital and Health Statistics*.

Summation Notation

In statistics, as in algebra, letters such as x, y, and z are used to denote variables. So, for instance, in a study of heights and weights of college students, we might let x denote the variable "height" and y denote the variable "weight."

We can often use notation for variables, along with other mathematical notations, to express statistics definitions and formulas concisely. Of particular importance, in this regard, is *summation notation*.

EXAMPLE 3.5 Introducing Summation Notation

Exam Scores The exam scores for the student in Example 3.4(a) are 88, 75, 95, and 100.

a. Use mathematical notation to represent the individual exam scores.

b. Use summation notation to express the sum of the four exam scores.

Solution

Let x denote the variable "exam score."

a. We use the symbol x_i (read as "x sub i") to represent the ith observation of the variable x. Thus, for the exam scores,

$$x_1 = \text{score on Exam 1} = 88;$$
$$x_2 = \text{score on Exam 2} = 75;$$
$$x_3 = \text{score on Exam 3} = 95;$$
$$x_4 = \text{score on Exam 4} = 100.$$

More simply, we can just write $x_1 = 88$, $x_2 = 75$, $x_3 = 95$, and $x_4 = 100$. The numbers 1, 2, 3, and 4 written below the xs are called **subscripts.** Subscripts do not necessarily indicate order but, rather, provide a way of keeping the observations distinct.

b. We can use the notation in part (a) to write the sum of the exam scores as

$$x_1 + x_2 + x_3 + x_4.$$

Summation notation, which uses the uppercase Greek letter Σ (sigma), provides a shorthand description for that sum. The letter Σ corresponds to the uppercase English letter S and is used here as an abbreviation for the phrase "the sum of." So, in place of $x_1 + x_2 + x_3 + x_4$, we can use **summation notation,** Σx_i, read as "summation x sub i" or "the sum of the observations of the variable x." For the exam-score data,

$$\Sigma x_i = x_1 + x_2 + x_3 + x_4 = 88 + 75 + 95 + 100 = 358.$$

Interpretation The sum of the student's four exam scores is 358 points.

Note the following about summation notation:

- When no confusion can arise, we sometimes write Σx_i even more simply as Σx.
- For clarity, we sometimes use **indices** to write Σx_i as $\sum_{i=1}^{n} x_i$, which is read as "summation x sub i from i equals 1 to n," where n stands for the number of observations.

The Sample Mean

In the remainder of this section and in Sections 3.2 and 3.3, we concentrate on descriptive measures of samples. In Section 3.4, we discuss descriptive measures of populations and their relationship to descriptive measures of samples.

Recall that values of a variable for a sample from a population are called *sample data*. The mean of sample data is called a **sample mean**. The symbol used for a sample mean is a bar over the letter representing the variable. So, for a variable x, we denote a sample mean as $\bar{x}$, read as "x bar." If we also use the letter n to denote the **sample size** or, equivalently, the number of observations, we can express the definition of a sample mean concisely.

DEFINITION 3.4

What Does It Mean?

A sample mean is the arithmetic average (mean) of sample data.

Sample Mean

For a variable x, the mean of the observations for a sample is called a **sample mean** and is denoted $\bar{x}$. Symbolically,

$$\bar{x} = \frac{\Sigma x_i}{n},$$

where n is the sample size.

EXAMPLE 3.6 The Sample Mean

Children of Diabetic Mothers The paper "Correlations Between the Intrauterine Metabolic Environment and Blood Pressure in Adolescent Offspring of Diabetic Mothers" (*Journal of Pediatrics*, Vol. 136, Issue 5, pp. 587–592) by N. Cho et al. presented findings of research on children of diabetic mothers. Past studies showed that maternal diabetes results in obesity, blood pressure, and glucose tolerance complications in the offspring.

Table 3.5 presents the arterial blood pressures, in millimeters of mercury (mm Hg), for a sample of 16 children of diabetic mothers. Determine the sample mean of these arterial blood pressures.

TABLE 3.5
Arterial blood pressures of 16 children of diabetic mothers

81.6	84.1	87.6	82.8
82.0	88.9	86.7	96.4
84.6	104.9	90.8	94.0
69.4	78.9	75.2	91.0

Solution Let x denote the variable "arterial blood pressure." We want to find the mean, $\bar{x}$, of the 16 observations of x shown in Table 3.5. The sum of those

observations is $\Sigma x_i = 1378.9$. The sample size (or number of observations) is 16, so $n = 16$. Thus,

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{1378.9}{16} = 86.18.$$

Exercise 3.31
on page 99

Interpretation The mean arterial blood pressure of the sample of 16 children of diabetic mothers is 86.18 mm Hg.

THE TECHNOLOGY CENTER

All statistical technologies have programs that automatically compute descriptive measures. In this subsection, we present output and step-by-step instructions for such programs.

EXAMPLE 3.7 Using Technology to Obtain Descriptive Measures

Weekly Salaries Use Minitab, Excel, or the TI-83/84 Plus to find the mean and median of the salary data for Data Set I, displayed in Table 3.1 on page 91.

Solution We applied the descriptive-measures programs to the data, resulting in Output 3.1. Steps for generating that output are presented in Instructions 3.1.

OUTPUT 3.1 Descriptive measures for Data Set I

MINITAB

Descriptive Statistics: SALARY

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
SALARY	13	0	483.8	73.7	265.8	300.0	300.0	400.0	625.0	1050.0

EXCEL

Summary of SALARY

Count	13
Mean	483.846
Median	400
Std Dev	265.786
Variance	70642.308
Range	750
Min	300
Max	1050
IQR	237.5
25th%	300
75th%	537.5

TI-83/84 PLUS

```
1-Var Stats
 x̄=483.8461538
 Σx=6290
 Σx²=3891100
 Sx=265.7862067
 σx=255.3591276
↓n=13
```

```
1-Var Stats
↑n=13
 minX=300
 Q1=300
 Med=400
 Q3=625
 maxX=1050
```

As shown in Output 3.1, the mean and median of the salary data for Data Set I are 483.8 (to one decimal place) and 400, respectively.

INSTRUCTIONS 3.1 Steps for generating Output 3.1

MINITAB

1. Store the data from Table 3.1 in a column named SALARY
2. Choose **Stat ➤ Basic Statistics ➤ Display Descriptive Statistics...**
3. Specify SALARY in the **Variables** text box
4. Click **OK**

EXCEL

1. Store the data from Table 3.1 in a range named SALARY
2. Choose **DDXL ➤ Summaries**
3. Select **Summary of One Variable** from the **Function type** drop-down box
4. Specify SALARY in the **Quantitative Variable** text box
5. Click **OK**

TI-83/84 PLUS

1. Store the data from Table 3.1 in a list named SAL
2. Press **STAT**
3. Arrow over to **CALC**
4. Press **1**
5. Press **2nd ➤ LIST**
6. Arrow down to SAL and press **ENTER** twice

Exercises 3.1

Understanding the Concepts and Skills

3.1 Explain in detail the purpose of a measure of center.

3.2 Name and describe the three most important measures of center.

3.3 Of the mean, median, and mode, which is the only one appropriate for use with qualitative data?

3.4 True or false: The mean, median, and mode can all be used with quantitative data. Explain your answer.

3.5 Consider the data set 1, 2, 3, 4, 5, 6, 7, 8, 9.
a. Obtain the mean and median of the data.
b. Replace the 9 in the data set by 99 and again compute the mean and median. Decide which measure of center works better here, and explain your answer.
c. For the data set in part (b), the mean is neither central nor typical for the data. The lack of what property of the mean accounts for this result?

3.6 Complete the following statement: A descriptive measure is *resistant* if ....

3.7 Floor Space. The U.S. Census Bureau compiles information on new, privately owned single-family houses. According to the document *Characteristics of New Housing*, in 2006 the mean floor space of such homes was 2469 sq ft and the median was 2248 sq ft. Which measure of center do you think is more appropriate? Justify your answer.

3.8 Net Worth. The Board of Governors of the Federal Reserve System publishes information on family net worth in the *Survey of Consumer Finances*. In 2004, the mean net worth of families in the United States was $448.2 thousand and the median net worth was $93.1 thousand. Which measure of center do you think is more appropriate? Explain your answer.

*In Exercises **3.9–3.14**, we have provided simple data sets for you to practice the basics of finding measures of center. For each data set, determine the*
a. mean. *b. median.* *c. mode(s).*

3.9 4, 0, 5

3.10 3, 5, 7

3.11 1, 2, 4, 4

3.12 2, 5, 0, −1

3.13 1, 9, 8, 4, 3

3.14 4, 2, 0, 2, 2

*In Exercises **3.15–3.22**, find the*
a. mean. *b. median.* *c. mode(s).*
For the mean and the median, round each answer to one more decimal place than that used for the observations.

3.15 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

3.16 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 shows 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

3.17 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

3.18 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

3.19 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

3.20 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

3.21 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

3.22 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 (hp) for each of the seven routers tested.

1.75	2.25	2.25	2.25	1.75	2.00	1.50

3.23 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 mean, median, and mode of these data.
b. Which measure of center do you think works best here? Explain your answer.

3.24 Monthly Motorcycle Casualties. The *Scottish Executive*, Analytical Services Division Transport Statistics, compiles data on motorcycle casualties. During one year, monthly casualties 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. Find the mean, median, and mode of the number of motorcycle casualties for built-up roads.
b. Find the mean, median, and mode of the number of motorcycle casualties for non–built-up roads.
c. If you had a list of only the month of each casualty, what month would be the modal month for each type of road?

3.25 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. Find the mean and median of the number of accidents for built-up roads.
b. Find the mean and median of the number of accidents for non–built-up roads.
c. If you had a list of only the day of the week for each accident, what day would be the modal day for each type of road?
d. What might explain the difference in the modal days for the two types of roads?

3.26 Explain what each symbol represents.
a. Σ **b.** n **c.** $\bar{x}$

3.27 For a particular population, is the population mean a variable? What about a sample mean?

3.28 Consider these sample data: $x_1 = 1$, $x_2 = 7$, $x_3 = 4$, $x_4 = 5$, $x_5 = 10$.
a. Find n. **b.** Compute Σx_i. **c.** Determine $\bar{x}$.

3.29 Consider these sample data: $x_1 = 12$, $x_2 = 8$, $x_3 = 9$, $x_4 = 17$.
a. Find n. **b.** Compute Σx_i. **c.** Determine $\bar{x}$.

*In each of Exercises **3.30–3.33**,*
a. find n.
b. compute Σx_i.
c. determine the sample mean. Round your answer to one more decimal place than that used for the observations.

3.30 Honeymoons. Popular destinations for the newlyweds of today are the Caribbean and Hawaii. According to a recent *American Wedding Study* by the Conde Nast Bridal Group, a honeymoon, on average, lasts 9.4 days and costs $5111. A sample of 12 newlyweds reported the following lengths of stay of their honeymoons.

5	14	7	10	6	8
12	9	10	9	7	11

3.31 Sleep. In 1908, W. S. Gosset published the article "The Probable Error of a Mean" (*Biometrika*, Vol. 6, pp. 1–25). In this pioneering paper, written under the pseudonym "Student," Gosset introduced what later became known as Student's t-distribution, which we discuss in a later chapter. Gosset used the following data set, which shows the additional sleep in hours obtained by a sample of 10 patients given laevohysocyamine hydrobromide.

1.9	0.8	1.1	0.1	−0.1
4.4	5.5	1.6	4.6	3.4

3.32 Pesticides in Pakistan. Pesticides are chemicals often used in agriculture to control pests. In Pakistan, 70% of the population depends on agriculture, and pesticide use there has increased rapidly. In the article, "Monitoring Pesticide Residues in Fresh Fruits Marketed in Peshawar, Pakistan" (*American Laboratory*, Vol. 37, No. 7, pp. 22–24), J. Shah et al. sampled the most commonly used fruit in Pakistan and analyzed the pesticide residues in the fruit. The amounts, in mg/kg, of the pesticide Dichlorovos for a sample of apples, guavas, and mangos were as follows.

0.2	1.6	4.0	5.4	5.7	11.4
0.2	3.4	2.4	6.6	4.2	2.7

3.33 U.S. Supreme Court Justices. From *Wikipedia*, we found that the ages of the justices of the U.S. Supreme Court, as of October 29, 2008, are as follows, in years.

53	88	72	72	69
60	75	70	58	

*In each of Exercises **3.34–3.41**,*
a. determine the mode of the data.
b. decide whether it would be appropriate to use either the mean or the median as a measure of center. Explain your answer.

3.34 Top Broadcast Shows. The networks for the top 20 television shows, as determined by the *Nielsen Ratings* for the week ending October 26, 2008, are shown in the following table.

CBS	ABC	CBS	ABC	ABC
Fox	CBS	CBS	Fox	CBS
ABC	CBS	CBS	CBS	Fox
Fox	Fox	CBS	Fox	ABC

3.35 NCAA Wrestling Champs. From NCAA.com—the official Web site for NCAA sports—we obtained the National Collegiate Athletic Association wrestling champions for the years 1984–2008. They are displayed in the following table.

Year	Champion	Year	Champion
1984	Iowa	1997	Iowa
1985	Iowa	1998	Iowa
1986	Iowa	1999	Iowa
1987	Iowa St.	2000	Iowa
1988	Arizona St.	2001	Minnesota
1989	Oklahoma St.	2002	Minnesota
1990	Oklahoma St.	2003	Oklahoma St.
1991	Iowa	2004	Oklahoma St.
1992	Iowa	2005	Oklahoma St.
1993	Iowa	2006	Oklahoma St.
1994	Oklahoma St.	2007	Minnesota
1995	Iowa	2008	Iowa
1996	Iowa		

3.36 Road Rage. The report *Controlling Road Rage: A Literature Review and Pilot Study* was prepared for the AAA Foundation for Traffic Safety by D. Rathbone and J. Huckabee. The authors discussed the results of a literature review and pilot study on how to prevent aggressive driving and road rage. As described in the study, *road rage* is criminal behavior by motorists characterized by uncontrolled anger that results in violence or threatened violence on the road. One of the goals of the study was to determine when road rage occurs most often. The days on which 69 road rage incidents occurred are presented in the following table.

F	F	Tu	Tu	F	Su	F	F	Tu	F
Tu	Sa	Sa	F	Sa	Tu	W	W	Th	Th
Th	Sa	M	Tu	Th	Su	W	Th	W	Tu
Tu	F	Th	Th	F	W	F	Th	F	Sa
F	W	W	F	Tu	W	W	Th	M	M
F	Su	Tu	F	W	Su	W	Th	M	Tu
F	W	Th	M	Su	Sa	Sa	F	F	

3.37 U.S. Supreme Court Justices. From *Wikipedia*, we found that the law schools of the justices of the U.S. Supreme Court, as of October 29, 2008, are as follows.

Harvard	Northwestern	Harvard
Harvard	Harvard	Yale
Columbia	Harvard	Yale

3.38 Robbery Locations. The Department of Justice and the Federal Bureau of Investigation publish a compilation on crime statistics for the United States in *Crime in the United States*. The following table provides a frequency distribution for robbery type during a one-year period.

Robbery type	Frequency
Street/highway	179,296
Commercial house	60,493
Gas or service station	11,362
Convenience store	25,774
Residence	56,641
Bank	9,504
Miscellaneous	70,333

3.39 Freshmen Politics. The Higher Education Research Institute of the University of California, Los Angeles, publishes information on characteristics of incoming college freshmen in *The American Freshman*. In 2000, 27.7% of incoming freshmen characterized their political views as liberal, 51.9% as moderate, and 20.4% as conservative. For this year, a random sample of 500 incoming college freshmen yielded the following frequency distribution for political views.

Political view	Frequency
Liberal	160
Moderate	246
Conservative	94

3.40 Medical School Faculty. The Women Physicians Congress compiles data on medical school faculty and publishes the results in *AAMC Faculty Roster*. The following table presents a frequency distribution of rank for medical school faculty during one year.

Rank	Frequency
Professor	24,418
Associate professor	21,732
Assistant professor	40,379
Instructor	10,960
Other	1,504

3.41 An Edge in Roulette? An American roulette wheel contains 18 red numbers, 18 black numbers, and 2 green numbers. The following table shows the frequency with which the ball landed on each color in 200 trials.

Number	Red	Black	Green
Frequency	88	102	10

Working with Large Data Sets

*In each of Exercises **3.42–3.50**, use the technology of your choice to obtain the measures of center that are appropriate from among the mean, median, and mode. Discuss your results and decide which measure of center is most appropriate. Provide a reason for your answer. Note: If an exercise contains more than one data set, perform the aforementioned tasks for each data set.*

3.42 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.

3.43 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

3.44 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 researchers' survey results, are provided on the WeissStats CD.

3.45 Ballot Preferences. In Issue 338 of the *Amstat News*, then-president of the American Statistical Association Fritz 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.

3.46 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.

3.47 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.

3.48 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.

3.49 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.

3.50 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 $1, 2, 3, \ldots$ 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