Everywhere around you, information is collected and stored. For example, the sets of data below list (in chronological order) sale prices for homes in two neighborhoods in Newark, New Jersey (from http://www.homes.com/Sold-Homes-Prices).

Zip Code 07102:
$270,000  $150,000  $288,000  $195,000  $430,000
$432,000  $305,000  $277,000  $172,000  $350,000
$375,000  $287,000  $278,000  $150,000  $180,000
$234,000  $157,000  $260,000  $110,000  $290,000

Zip Code 07104:
$300,000  $330,000  $212,000  $260,000  $63,000
$275,000  $132,000  $185,000  $136,000  $210,000
$310,000  $95,000   $32,000   $152,000  $200,000
$350,000  $41,000   $297,000  $237,000  $250,000

**Why might you** be interested in these numbers? Perhaps you would like to know what you might expect to pay for a house. Perhaps you would like to know in which neighborhood you could buy a house more economically. Perhaps you own a house and want to appeal your property tax appraisal. In any case, a large collection of data, such as the housing prices for a large city, isn’t very useful until it is organized and analyzed. This chapter introduces you to several techniques for organizing and analyzing data.

According to the American Statistical Association’s home page,

Statisticians contribute to scientific enquiry by applying their mathematical and statistical knowledge to the design of surveys and experiments; the collection, processing, and analysis of data; and the interpretation of the results. Statisticians may apply their knowledge of statistical methods to a variety of subject areas, such as biology, economics, engineering, medicine, public health, psychology, marketing, education, and sports. Many economic, social, political, and military decisions cannot be made without statistical techniques, such as the design of experiments to gain federal approval of a newly manufactured drug.

---

10.1 Introduction to Statistics
10.2 Descriptive Statistics: Averages
10.3 Variation and Summary
10.4 Linear Regression: Finding Equations from Data

**CHAPTER 10 SUMMARY**
**CHAPTER 10 REVIEW**
**EXERCISES**
**CHAPTER 10 TEST**
## 10.1 Introduction to Statistics

**OBJECTIVES**
- Identify the terms statistics, population, census, sample, population parameters, and sample statistics.
- Identify the steps in a statistical study.
- Identify a few methods of sampling.
- Identify sources of bias.

### WARM-UP
1. Arrange the sale prices for zip code 07102 in order from lowest to highest.
2. Arrange the sale prices for zip code 07104 in order from lowest to highest.

**YOU MIGHT HAVE HEARD** questions like these in the media:
- Is a new drug a cure for cancer?
- What is an estimated range of prices for a house to be sold in zip code 07102?
- What is the average number of keys carried by your classmates? What might be concluded about the number of keys?
- What is the average neck size of a grizzly bear in Yellowstone National Park?
- What was the 2010 population in the United States?

Or maybe you heard statements like these:
- 4 out of 5 dentists prefer a particular brand of toothpaste.
- The average salary of an NBA player in the 2007–2008 season was $5.356 million per year.
- Men who carry their billfold in a hip pocket are more likely to develop certain back problems.

Whether reasonable or not, these are examples of statistics.

**BASIC DEFINITIONS**

Statistics is the science of collecting, organizing, and interpreting data. A statistic can also be a number that summarizes a data set.

Like all sciences, statistics has its own set of terms and processes.

---

**Definitions of Statistical Terms**

**Population**: set (of people or things) being studied

**Sample**: a portion (usually small) of the population

**Census**: collection of information about every item in a population set. Most of the time it is impractical or impossible to do a census—we must select a sample.

**Population parameters**: specific characteristics of the population that are being studied
EXAMPLE 1  Identifying terms in a statistical study  The National Highway Traffic Safety Administration crashes 25 2010 Ford Escape SUVs at speeds of less than 5 mph and then determines the average cost of repair. The SUVs are taken from six different Ford manufacturing plants throughout the country. Name the population, population parameter, sample, and sample statistics for the study. Is this an experimental or observational study?

Solution  Population: all 2010 Ford Escape SUVs  Population parameter: the average cost of repair for a 5-mph crash for this SUV  Sample: the 25 SUVs selected from the six different plants  Sample statistic: the average repair cost for these 25 SUVs  The study is experimental.

PRACTICE 1  Return to the lists of questions and statements below the Warm-up.

a. Testing a new cancer drug requires a (an) __________ study.

b. If you wanted to determine the average prices of all homes in Newark, the numbers in the chapter opener would allow you to calculate a (an) __________. All houses sold in Newark might be a (an) __________. The average sales price is the __________ parameter for the houses in the chapter opener.

c. Taking a count of the number of people in the United States is an attempt to take a (an) __________.

d. To determine the neck sizes of grizzly bears in Yellowstone National Park, a (an) __________ study would need to be conducted.

e. What are the populations in the three media statements on page 542?

PRACTICE 2  Joe Candidate decides to quit his day job and run for governor. He conducts a name-recognition poll of 250 potential voters and finds that his name is recognized by 10% of those voters. Identify the population, population parameter, sample, and sample statistic. Is this an experimental or observational study?

• STEPS IN A STATISTICAL STUDY

A well-designed statistical study permits someone else with the same observations or experiments to replicate your results. The first column of Table 1 lists the traditional steps in a statistical study. The problem-solving skills we have used since Section 1.1 are listed in the second column. This arrangement permits you to compare steps and see the logic behind the process of conducting an observational or experimental study.
TABLE 1  Comparing a Statistical Study and Problem Solving

<table>
<thead>
<tr>
<th>Steps in a Statistical Study</th>
<th>Steps in Problem Solving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pose the question to be studied and identify the population.</td>
<td>Understand the problem.</td>
</tr>
<tr>
<td>Choose a sample and collect raw data.</td>
<td>Make a plan.</td>
</tr>
<tr>
<td>Create the sample statistics.</td>
<td>Carry out the plan.</td>
</tr>
<tr>
<td>Infer* characteristics of the population.</td>
<td>Check your solution.</td>
</tr>
<tr>
<td>Draw conclusions.</td>
<td>Extend to other settings.</td>
</tr>
</tbody>
</table>

*Infer: Arrive at a decision or opinion by reasoning from known facts or evidence (Webster's New World College Dictionary).

**EXAMPLE 2**  Identifying steps in a statistical study  You are facilities manager for your campus. Space for a new fast-food provider will be opening up in your student center at the end of the year. Suppose you wish to study whether students, staff, and faculty at your college prefer fast food from company A or from company B. What steps might you use?

**Solution**  Your population is people at your college. Assuming the population is large, you should choose a sample and collect the raw data from that sample (having members of the sample answer the question “Do you prefer company A or company B?”). From the data you might calculate a sample statistic such as the percentage of people who prefer company A over company B. Suppose you find that 56% of the people in the sample prefer company A. Although this does not mean that 56% of the entire population prefers company A, you can probably make some assumptions about how many prefer company A.

**OBTAINING A SAMPLE**

Collecting a good, representative sample is extremely important in every statistical study. Here are questions you should ask about a sample.

- Is the sample a random sample, in which each member of the population has an equal chance of being selected? If samples of a given size are being chosen, does each sample have an equal chance of being selected?
- Is a systematic sample appropriate? Does selecting every 5th, 10th, or nth member of the population give a random or representative sample?
- Is sampling from groups appropriate? Are different groups represented within the sample in the same proportion as in the population? Sampling from groups in statistics is called stratified sampling.
- Is the sample selected strictly because of convenience—a convenience sample? (Look for examples in this section of why convenience is not a recommended method!)

**EXAMPLE 3**  Obtaining a good sample  As campus facilities manager (see Example 2), you talk with a statistics instructor in the mathematics department before selecting your sample. What might the statistics instructor say about each of the following choices of sampling techniques?

a. random sample  
b. systematic sample  
c. stratified sampling  
d. convenience sample
**Solution**

a. A random sample from all people on campus would be a good way to reach those who currently use campus food services and those who might be attracted to a new service.

b. She offers to have her statistics class do a survey of current users of food services. She will have them select every \( n \)th person in the food line at certain times of day, remembering that lunch time is the time of heaviest demand. Her selection will also consider the MWF and TuTh class schedules.

c. The campus has two groups: 800 faculty and staff and 8000 students. Although faculty and staff are likely to be on campus five days a week, students tend to make up a larger share of the current customers. She suggests that your sample represent the two groups in a ratio of 1 to 10.

d. She cautions you about looking for an easy way to get opinions, such as asking attendees at the next Head of Services meeting with the college vice president. This group regularly has luncheon meetings off campus and is seldom seen in the cafeteria for any reason.

**PRACTICE 3**

What method of sampling is described?

a. In order to complete a student survey for accreditation,* your college uses a computer program to randomly choose 100 student identification numbers. These students are asked to answer the survey questions.

b. The *New York Times* summarized an investigation by the British General Medical Council (BGMC) of a researcher who drew blood samples from children at his son’s birthday party and used them in his research. The BGMC withdrew the researcher’s license to practice medicine in England.†

c. You are doing a survey to investigate racial prejudice. You classify the United States population into five different racial groups: White, Hispanic, African American, American Indian, and Asian. You take a random sample from each group, using the same proportions as in the population, and ask the people in your samples to complete the survey.

**PRACTICE 4**

Return to the home sale prices in the chapter opener.

a. Choose every fifth sale price from the 07102 zip code and find the sum. Divide by 4 to find an average. What type of sampling is this?

b. The first four of the five digits selected randomly by the calculator in the Graphing Calculator Technique Box on page 547 are 7, 16, 11, and 9. Let these correspond to the 7th, 16th, 11th, and 9th prices in the chapter opener list for the 07102 zip code. Find the sum, and divide by 4 to find an average. What type of sampling is this?

c. How comfortable are you with using these averages in describing sale prices for that zip code?

**AVOIDING BIAS**

Misuse of the statistical process or its results has caused statistics to get a bad reputation. Bias, a slanting of information or partiality at any step in the statistical study, can lead to bad results and, hence, a bad reputation for the science of statistics.

---

*The goal of accreditation is to ensure that education provided by institutions of higher education meets acceptable levels of quality (http://www.ope.ed.gov/accreditation).
Table 2 contains several questions to ask at each step in a statistical study in order to avoid bias.

**TABLE 2** Questions to Ask about Statistical Studies to Avoid Bias

<table>
<thead>
<tr>
<th>Steps in a Statistical Study</th>
<th>Questions to Help Prevent Bias</th>
</tr>
</thead>
</table>
| Pose the question to be studied and identify the population.| • Is the question well defined?  
• Is the question measurable?  
• Is the population appropriate?  
• Is the type of study appropriate to the question?             |
| Choose a sample and collect raw data.                       | • How is the sample to be obtained?  
• Is the sample random, systematic (if appropriate), or grouped (if appropriate)?  
• Was a convenience sample avoided?  
• Are participants casual volunteers or otherwise not representative of the population? Will participants reply to a survey?  
• Does the wording and/or setting of the survey bias the result? |
| Create the sample statistics.                               | • Are statistics selected appropriate?  
• Are statistics used correctly?                                 |
| Infer characteristics of the population.                   | • Is the interpretation consistent with the results?                                           |
| Draw conclusions.                                           | • Are the results presented fairly?  
• Does the conclusion make sense?  
• Are there alternative reasons for the results?  
• Does the researcher have a conflict of interest?  
• What is the source of the study?  
• Who funded it?  
• Who benefits from it?  
• Who might benefit from hiding the results?                   |

**PRACTICE 5**

What questions should be (or should have been) asked in these settings?

a. The *New York Times* reported findings by the British General Medical Council that a researcher studying vaccines received part of his funding from lawyers for parents seeking to sue vaccine makers for damages.*

b. The *New York Times* report in part a also indicated that this researcher had a patent on an alternative vaccine that could be used if his research suggested a link between autism and the standard vaccine, which has a long and well-documented history of use.

c. Two Australian scientists, Barry Marshall and Robin Warren, were frustrated because their 1982 research conclusions that common forms of ulcers were caused by the *H. pylori* bacteria (and hence curable with an antibiotic) were not published in the United States and were generally ignored elsewhere. They received a Nobel Prize in 2005.

d. In conducting a survey, Taxpayer’s Revolt, a group supporting tax decreases, asked, “Do you favor balancing the budget by raising taxes or eliminating wasteful government programs?”

e. ADT security systems advertised, “When you go on vacation, burglars go to work.” The ad stated that, according to FBI statistics, over 26% of home burglaries take place between Memorial Day and Labor Day.

f. To determine the effectiveness of their program, a college English department mailed 1000 recent graduates a survey.

Formally, statistical bias is any systematic error contributing to the difference between the statistical values in a population and a sample drawn from it (Webster’s New World College Dictionary).

Graphing Calculator Technique: Obtaining a Random Sample

Using the random integer, or randInt, option on a calculator is an easy way to select a random sample. In the Activity on page 549, you have 20 individuals. Suppose you would like to select a sample of size 5. To create a random sample, you assign a number, 1 to 20, to each individual. On the calculator, press [MATH]. Then press either the left or the right cursor arrow to move to the PRB (probability) option. The menu changes for each of the options across the top. The menu for PRB is shown in Figure 2. To select random integer, press 5. Your screen should then match Figure 3.

![Figure 2](probability menu)

Enter your starting and ending numbers, separated by a comma, and then enter the number of random integers requested: 1, 20, 5. Finish with a closing parenthesis, as shown in Figure 4. Press [ENTER] to randomly generate five numbers.

![Figure 4](requesting five integers on [1, 20])

![Figure 5](five random integers)

Figure 5 shows one set of random integers on the interval [1, 20]. The numbers indicate that your sample is the individuals assigned numbers 7, 16, 11, 8, and 13.
9, and 13. Pressing \( \text{ENTER} \) again gives more samples. You do not need to reenter the command \texttt{randInt} (1, 20, 5). Note: It is possible to get the same number selected twice—for example, 8, 1, 11, 11, 15. In case of duplicates, the usual technique is to ignore the sample and simply press \( \text{ENTER} \) again for another sample.

**Answer Box**

**Warm-up:**

1. Zip code 07102: 110,000, 150,000, 150,000, 157,000, 172,000, 180,000, 195,000, 234,000, 260,000, 270,000, 277,000, 278,000, 287,000, 288,000, 290,000, 305,000, 350,000, 375,000, 430,000, 432,000

2. Zip code 07104: 32,000, 41,000, 63,000, 95,000, 132,000, 136,000, 152,000, 185,000, 200,000, 210,000, 212,000, 237,000, 250,000, 260,000, 275,000, 297,000, 300,000, 310,000, 330,000, 350,000

**Practice 1:**

a. experimental  

b. sample statistic, population, population  

c. census  

d. observational (although dangerously hands-on unless you gather data at a museum)  

e. all dentists, all NBA players in the 2007–2008 season, all men

**Practice 2:**

population: all potential voters in Joe Candidate’s state; population parameters: the percentage who recognize Joe Candidate; sample: the 250 potential voters selected; sample statistics: the 10% name recognition. The study is observational.

**Practice 3:**

a. random sampling  

b. convenience sampling  

c. grouped (stratified) sampling

**Practice 4:**

a. $1,250,000; $312,500; systematic sampling

b. $1,086,000; $271,500; random sampling

**Practice 5:**

a. Who paid for the research?  

b. Is there any conflict of interest?  

c. Who benefited from hiding the results? (Pharmaceutical companies eventually responded by selling “ulcer” medication without prescriptions.)  

d. Does the wording (wasteful) bias the results?  

e. Is interpretation consistent with results? (The number of days between Memorial Day and Labor Day is 98, or 26.8% of the year. If burglaries are random, we would expect 26% to happen between those dates.)  

f. Will participants reply to a survey? (It is optional, resulting in self-selection bias.)

**Reading Questions**

1. Statistics is both the ______________ of collecting, organizing, and interpreting data and the numbers, calculations, or observations that ______________ the data.

2. A (An) ______________ is a selection made from the population.

3. What are the five steps in a statistical study?

4. Samples should be grouped (or stratified) in the same ______________ as the population (as appropriate).

5. Statistical ______________ is any systematic error contributing to the difference between the statistical values in a population and a sample drawn from it.
ACTIVITY

Classroom Keys. One day your mathematics instructor gives an individually numbered card to each of the 20 students in the class. She asks each student to count the number of keys in his or her purse or pocket that day. She summarizes the information from the cards in Table 3.

<table>
<thead>
<tr>
<th>Student Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Keys</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>12</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Number</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Keys</td>
<td>10</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

1. Determine the population for this study.

2. Calculate the average number keys students had with them that day. (To find the average; add all the numbers of keys and divide by the number of students.) This is your population parameter.

3. Now take a simple random sample. Use the technique described in the Graphing Calculator Technique box to find random integers on the interval [1, 20]. Collect five different random samples of size 6. Remember that each random integer represents a student number, not a number of keys. For each random sample, calculate the average number of keys held by students in that sample. Each average is a sample statistic.

4. Compare each sample statistic to the population parameter. Is it identical, larger, or smaller?

5. Calculate the percentage error between each sample statistic and the population parameter. (Subtract the sample statistic from the population parameter and divide by the population parameter. Mentally use absolute value in the subtraction, as we are not interested in the sign on the error.)

This activity illustrates the concept of sampling variability. Sampling variability happens even when you take a random sample correctly—that is, without bias. Your sample result will most likely differ from the population parameter, sometimes significantly.

10.1 Exercises

In Exercises 1 to 5, indicate whether the given study is an observational study or an experimental study. Name the population(s) being studied, the sample(s), and the statistic given.

1. A Gallup poll done for CNN surveyed 2010 randomly selected Americans and found that 80% of Americans think we are less civil than we were 10 years ago and 67% think we are more likely to use vulgar language than we were 10 years ago.

2. A survey of 235,812 first-year college students revealed that 32.4% of these students had an A average in high school.

3. A psychologist wants to determine if the color of our surroundings affects our concentration. To do so, he randomly selects a sample of individuals and first asks them to complete a puzzle in a white room and then asks them to complete a similar puzzle in a bright pink room. He finds that 75% of the individuals could complete the task more quickly in the white room.

Circled numbers are core exercises.
4. To measure the effect of air pollution on sparrow populations, a scientist finds historical Christmas bird count data for sparrows in polluted areas and compares these data to Christmas bird count data for the same species of sparrow in unpolluted areas. She finds that the number of sparrows in the unpolluted areas exceeds the number in polluted areas by 453.

5. A college instructor wants to know if playing music during an exam will improve test scores. He randomly divides his class into two groups and gives each group the same test. The first group takes the test in the usual classroom with no music. The second group takes the test in a second classroom where classical music is being played. The instructor discovers that the students who were in the classroom with music scored, on average, 15 points higher on the exam than those who were in the regular classroom.

In Exercises 6 to 11, identify the sampling method used and indicate whether it is a valid method.

6. You are doing a survey to investigate racial prejudice. You classify the U.S. population into five different racial groups: White, Hispanic, African American, American Indian, and Asian. You take a random sample from each group and use these people as the sample for your study.

7. You manufacture light bulbs and are interested in how many bulbs in each batch of 1000 bulbs are defective. You test each 10th bulb off the assembly line to determine if it is defective.

8. To determine opinions on the state legislature’s raising of tuition at public colleges in your state, you ask everyone in the food court at the mall on a given Friday afternoon.

9. A university is interested in students’ opinions about effective use of Student Activities funds. The student population is divided into undergraduate students and graduate students, and a random sample is taken from each group to construct the entire sample.

10. You design two T-shirts and are interested in whether people would buy the shirts. You go to the park and ask people seated in the playground area for their opinions of the shirts.

11. To conduct a phone survey, a company has a computer randomly generate telephone numbers and uses the people who answer calls to these numbers as their sample.

In Exercises 12 and 13, identify and evaluate the following parts of a statistical study.

(a) What is the population for the study? What is the sample for the study?

(b) What are the raw data for the study?
(c) What is the sample statistic for the study?
(d) Analyze the sampling technique of the experimenter—do you believe the results will be accurate? Why or why not?

12. You wish to study the amount of time the 3500 students at your college spend exercising each day. You ask 100 students who live in the dorms to fill out a survey on daily exercise. Eighty-five students fill out and return the survey. Based on the results, you find that the average amount of daily exercise is 30 minutes. You conclude that, on average, the 3500 students at your college exercise about 30 minutes per day.

13. You wish to study the amount of time the 3000 students at your college watch TV each day. You ask 100 students who are enrolled in a Problem Solving course to fill out a survey on daily TV viewing time. Twenty-five students fill out and return the survey. Based on the results, you find that the average amount of TV watched per day is 3.25 hours. You conclude that, on average, the 3000 students at your college watch TV about 3.25 hours per day.

In Exercises 14 to 19, identify the bias in the study described.

14. A student researcher does a survey to determine if college professors or scuba divers are more “satisfied” with their lives.

15. The Newport Chronicle claims that pregnant mothers can increase their chances of having healthy babies by eating lobsters. The claim is based on a study showing that babies born to mothers with lobster in their diet have fewer health problems than babies born to mothers who don’t eat lobster.

16. For about 60 years, the Gallup organization has been polling people about their attendance at religious services. A recent poll of Roman Catholics in a county in Ohio showed that 51% said they had attended church the week before the poll. Yet actual attendance reported by churches was 24%.

17. Kiwi Brands, a shoe polish manufacturer, makes the following claim in an advertisement: “According to a nationwide survey of 250 hiring professionals, scuffed shoes was the most common reason for a male job seeker’s failure to make a good impression.”

18. In order to survey the entire student body at your college, you stand outside the cafeteria and interview every fifth student.

19. In November 1997, the leadership of Congress was considering whether to spend $30 million for a survey of every taxpayer. Response to the survey would be
voluntary. The Speaker of the House of Representatives was quoted in the New York Times as saying, “For less than 50 cents a return, we think every American who pays taxes deserves the right to tell the government how well it collects their money.” The Times noted that Congress had already paid for a professional poll of attitudes about the IRS, and that poll cost only $20,000.

20. A survey was conducted of patrons at a local ice cream shop on a Saturday afternoon to determine their favorite type of ice cream. Five categories were used: peppermint (P), strawberry (S), chocolate (C), vanilla (V), and Rocky Road (R). The survey results are given below:


a. What is the population of this study? What is the sample?
b. Calculate the percentage who prefer each type of ice cream.
c. Would you consider these results valid estimates of the population parameters for this population? Why or why not?

21. Return to the home sale prices in the chapter opener.
a. Choose every fifth sale price from the 07104 zip code and find the sum. Divide by 4 to find an average. What type of sampling is this?
b. The first four of the five digits selected randomly by the calculator in the Graphing Calculator Technique box on page 547 are 7, 16, 11, and 9. Let these correspond to the 7th, 16th, 11th, and 9th prices in the chapter opener list for the 07104 zip code. Find the sum and divide by 4 to find an average. What type of sampling is this?
c. How comfortable are you with these averages to describe the sale prices for that zip code?
d. Compare your results with those in Practice 4. What might you conclude in comparing the sale prices in the two zip codes? Based on your samples, how confident are you that one zip code has lower prices than the other? Do you think it is necessary to average all prices in each set?

22. A mathematics instructor gave a test in his course; the results appear below. The test was worth a total of 100 points. You may assume that each student in the course took the test.

| 74 | 65 | 98 | 34 | 92 | 85 | 69 |
| 76 | 80 | 55 | 81 | 72 | 78 | 60 |

a. Is this a census or a sample?
b. Calculate the average test score.
c. Select three random samples of size 4 and calculate the average test score for each sample.
d. Compare the average scores for the samples you chose in part c with the average obtained in part b by calculating the percentage error.

23. The high temperatures (in degrees Fahrenheit) at the Spokane, Washington, airport for 30 days in March 2009 were as follows:

41 44 45 38 35 34 38 35 30 20
24 34 41 38 47 43 43 43 48 57
49 41 43 43 47 41 51 52 49 45

a. Calculate the average high temperature.
b. Select a systematic sample by choosing every sixth entry, starting with the first entry. Calculate the average temperature of this sample.
c. Select a random sample of size 6 and calculate the average of this sample.
d. Calculate the percentage error between the sample statistic in part b and the average calculated in part a. Repeat for part c with part a.

24. Project: Self-Selection Bias. One source of bias in a sample is the self-selection bias that arises when participants volunteer. Find three statistical reports in the media that may be biased because of self-selection. Explain why these reports suffer from self-selection bias.

25. Project: Nonresponse Bias. Another source of bias, called nonresponse bias, occurs when a survey is distributed to a random sample and a small percentage of the sample respond. Find three statistical reports in the media that may be biased because of nonresponse.

26. Project: Presidential Election. The 1948 presidential election was incorrectly predicted in favor of Dewey instead of Truman. Research this election and find out why predictions were wrong.

27. Project: Security Requirement Frustration. An airline executive has hired your research firm to find out whether there has been an increase in frustration among air travelers due to the increased security requirements at airports.
a. Explain how to apply the steps of a statistical study to this setting.
b. Discuss pros and cons of various means of obtaining data.
c. Suggest possible measures of frustration that don’t require interaction with the passengers.
Descriptive Statistics: Averages

OBJECTIVES

- Create frequency tables.
- Calculate measures of central tendency: the mean, median, and mode.
- Suggest why one measure of central tendency might be better than others for a given setting.

WARM-UP

1. Solve for $s$: \( \frac{63 + 59 + 85 + s}{4} = 75 \).
2. Solve for $r$: \( \frac{0.8 + 1.4r}{4} = 0.90 \).

TO CONDUCT A RELIABLE STATISTICAL STUDY, you usually need to collect or use a large set of data. When it is not practical to list each item in a set, the data should be condensed into a more manageable form. This section describes several common ways of organizing and summarizing data to make them easier to use and interpret. We will discuss descriptive statistics, which provide ways to describe some of the basic features of a set of data.

ORGANIZING DATA

The easiest way to organize quantitative data is to sort the numbers into ascending (smallest to largest) or descending (largest to smallest) order.

EXAMPLE 1 Sorting data Use a calculator to sort into ascending order the house prices for zip code 07102 given in the chapter opener (page 541).

Solution Step 1: Enter the data. Press \( \text{STAT} \) to obtain the statistics menu (see Figure 6). Press 1 to choose edit (and enter data). Lists will appear, as in Figure 7. You might want to return to the original order later, so enter 1 to 20 in the first list, L1. If something is already in the list, erase it by pressing the up arrow, \( \text{CLEAR} \), and the down arrow. Enter the prices in L2. It is not necessary to enter all the zeros if you remember that the house prices are in thousands. Enter all 20 prices under L2. L1 and L2 should be the same length.
Step 2: Sort the data. Press \texttt{STAT} for the statistics menu (Figure 6) and choose 2 : \texttt{SortA}, to sort in ascending order. The sort command will appear on your home screen, as in Figure 8. Nothing will happen until you enter the lists you want sorted. The lists are above the number keys 1 to 6. Press \texttt{2ND} [L2] \( \rightarrow \) \texttt{2ND} [L1] \( \rightarrow \) \texttt{ENTER} to obtain the list in ascending order, as in Figure 9.

Because you might want the original order, you have also entered L1. Try undoing the sort by choosing \texttt{SortA} and entering \texttt{2ND} [L1] \( \rightarrow \) \texttt{ENTER}.

We can also organize data sets by placing the information in a table. A frequency table lists the number of times a number or item occurs in a data set.

Table 4 shows household income data for Miami-Dade County, as reported by the U.S. Census Bureau in 2000. The data are grouped in order to make the table easier to read and to emphasize the relative numbers of households at certain levels of income. This table also gives median and mean incomes.

<table>
<thead>
<tr>
<th>Total Household Income</th>
<th>Number of Households</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than $10,000</td>
<td>86,170</td>
</tr>
<tr>
<td>$10,000 to $14,999</td>
<td>57,928</td>
</tr>
<tr>
<td>$15,000 to $24,999</td>
<td>105,156</td>
</tr>
<tr>
<td>$25,000 to $34,999</td>
<td>95,305</td>
</tr>
<tr>
<td>$35,000 to $49,999</td>
<td>121,309</td>
</tr>
<tr>
<td>$50,000 to $74,999</td>
<td>139,697</td>
</tr>
<tr>
<td>$75,000 to $99,999</td>
<td>83,818</td>
</tr>
<tr>
<td>$100,000 to $149,999</td>
<td>79,235</td>
</tr>
<tr>
<td>$150,000 to $199,999</td>
<td>27,867</td>
</tr>
<tr>
<td>$200,000 or more</td>
<td>31,446</td>
</tr>
</tbody>
</table>

Total number of households: 827,931
Median household income: $42,969
Mean household income: $63,299

- AVERAGES, OR MEASURES OF CENTRAL TENDENCY

The household incomes in Table 4 were summarized by a median and a mean, which help us answer the question “What is the typical household income in Miami-Dade County?” This seems like a simple question, but what does “typical” really mean? Does “typical” mean the most common household income, the middle one of all the incomes listed, or something else? Statistics that tell us about the center or “typical” result are called measures of central tendency.
An average describes the center, or middle, of a set of numbers. In some settings, we think of average as being “normal.” In other settings, an average helps us compare an individual with a group or compare one group with another group. In geometry, the average is a description of middle.

We now define three important ways of describing the middle of a set of numbers—with the mean, the median, and the mode.

**Mean** Usually, people say they are finding the “average” when they add numbers together and divide by how many numbers were added. This average is called the mean or mean average, to distinguish it from the other forms of averages (see median and mode, below).

### Finding the Mean

The mean of a set of numbers is the sum of the numbers divided by the number of numbers in the set.

In formal statistics, the Greek letter sigma, $\Sigma$, is a short way to write the sum of a set of numbers. The mean is $\frac{\Sigma x_i}{n}$, where the $x_i$ are the numbers in the set and $n$ is the number of numbers in the set.

#### EXAMPLE 2

**Finding the mean: mean income** Suppose the incomes of five families are $14,000, 14,000, 16,000, 17,000, and 100,000. Find the mean and explain why the mean does not give a good description of this set of families.

**Solution**

\[
\frac{14,000 + 14,000 + 16,000 + 17,000 + 100,000}{5} = \frac{322,000}{5} = 64,400
\]

The mean appears to indicate that a typical family has an income well above $18,310, the 2009 poverty level for families of three persons. Yet in reality, four of the five families are below the poverty level.

In Example 2, the $100,000 is an **outlier**, a data point that is not close to the others in the set. Defining “close” requires still more statistics. However, if your income is $13,000, someone else’s income of $32,200 or $100,000 is not close.

**Median** A very high or low number, as in Example 2, can cause the mean not to be close to most of the numbers. Because the mean does not always give a good description, statisticians invented other averages, such as the median and the mode.

### Finding the Median

The median is found by selecting the middle number when the numbers are arranged in numerical order (from smallest to largest or largest to smallest). If there is no single middle number, the median is the mean of the two middle numbers.

#### EXAMPLE 3

**Finding the median: median income** Find the median of each set of incomes.

a. $14,000, 14,000, 16,000, 17,000, 100,000

b. $50,000, 20,000, 80,000, 16,000

**Solution**

a. The median of $14,000, 14,000, 16,000, 17,000, and 100,000 is $16,000, because the numbers are already arranged in order and $16,000 is the middle number.
b. The second set of numbers must be rearranged into ascending order: $8000, $16,000, $20,000, and $50,000. The set has no middle number. In this case, we find the mean of the two middle numbers and use it as the median. The median is

$$\frac{16,000 + 20,000}{2} = 18,000$$

As suggested by Example 3, the median is more descriptive than the mean when one number in the set of data is some distance from the rest of the numbers (an outlier). Table 4 shows a median household income less than the mean household income because there are a few households with large incomes.

**Mode** The mode is a third form of average. The mode is useful when the average needs to describe a most popular or most common item. In the set of numbers [2, 2, 2, 3, 4], the mode is 2 because 2 appears most often.

If data are not numerical (describing, for example, race, occupation, college major, or eye color), the mode must be used.

**Finding the Mode**

The mode of a set is the number or item that occurs most often.

Table 5 shows a tally by states of the age at which a person may obtain a licence for unrestricted operation of a private passenger car. Many states with an age of 18 allow the applicant to drive at a younger age if he or she has taken a driver’s education course.

TABLE 5 Age for unrestricted operation of private passenger cars, 1991

<table>
<thead>
<tr>
<th>Driver’s Age (years)</th>
<th>States (including the District of Columbia)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>II</td>
</tr>
<tr>
<td>16</td>
<td>III</td>
</tr>
<tr>
<td>16.5</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>III</td>
</tr>
<tr>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>


**Example 4**

**Finding averages from tables** Find the mean, median, and mode of the data in Table 5.

**Solution**

To find the mean age for Table 5, we add the 51 ages and divide the total by 51:

$$\frac{15(2) + 16(20) + 16.5(1) + 17(3) + 18(24) + 19(1)}{51}$$

Entering the entire expression in a calculator and pressing **Enter** will yield a result of approximately 17.03 years.

The tally places the data in order by age, so we count to the middle tally mark. The middle of the 51 tally marks is the 26th tally mark, because it has 25 marks before it and 25 after it. The 26th tally mark is the last mark for age 17, so age 17 is the median.
The mode for minimum driving age is 18. The number of states with a minimum driving age of 16 is almost the same as the number with a minimum driving age of 18. If the numbers were exactly the same, the data would be bimodal—that is, having two modes.

**Graphing Calculator Technique: Calculator Mean and Median for Table 5**

Press \[ \text{STAT} \] 1: Edit. Under L1, place the various ages. Under L2, place the number of states with each age. L2 now gives the frequency of each item in L1 (see Figure 10). Omitting L2 indicates that the set contains only one of each item in L1.

**Figure 10** Age and tally in list.

**Figure 11** MATH menu.

Press 2ND [QUIT].

To find the mean, use 2ND [L1] MATH and choose 3: \text{mean}. Then press \text{ENTER} (see Figure 11). When \text{mean} appears, press 2ND [L1] \text{ mean} [2ND [L2] \text{ ENTER}.

To find the median, use 2ND [L1] MATH and choose 4: \text{median}. Then press \text{ENTER}. When \text{median} appears, press 2ND [L1] \text{ median} [2ND [L2] \text{ ENTER}.

Figure 12 shows the results for both the mean and the median.

**Practice 1**

a. Find the mean, median, and mode of the numbers of barrels of oil used daily by the 216 oil-consuming countries in Table 6. Assume that the bottom 199 countries each use 0.5 million barrels per day.

b. Why are the mean and median so far apart?
### DESCRIPITIVE STATISTICS: AVERAGES

#### TABLE 6  World Oil Consumption, 2008

<table>
<thead>
<tr>
<th>Million of Barrels per Day</th>
<th>Countries</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;10</td>
<td>United States (19.5)</td>
</tr>
<tr>
<td>7–10</td>
<td>China (7.8)</td>
</tr>
<tr>
<td>4–7</td>
<td>Japan (4.8)</td>
</tr>
<tr>
<td>2–4</td>
<td>India (3.0), Russia (2.9), Germany (2.6), Brazil (2.5), Saudi Arabia (2.4), Canada (2.3), South Korea (2.2), Mexico (2.1)</td>
</tr>
<tr>
<td>1–2</td>
<td>France (2.0), Iran (1.7), United Kingdom (1.7), Italy (1.6), Spain (1.6), Indonesia (1.3)</td>
</tr>
<tr>
<td>&lt;1</td>
<td>199 countries</td>
</tr>
</tbody>
</table>


### EXAMPLE 5  Choosing a measure of center

State what measure of center you would use for the following data sets. Explain your reasoning.

**a.** Typical salaries in your home state

**b.** Water usage of various countries

**c.** Scores you received on four exams

**d.** The results of a survey on customer service that uses a scale known as a Likert scale: 3 for excellent, 2 for fair, 1 for poor

**e.** The hair color of students at your school

### Solution

**a.** You could use either the median or the mean. However, the mean would be affected by outliers and would most likely be higher than the median.

**b.** You could use either the median or the mean. However, the mean would be affected by outliers and would most likely be higher than the median.

**c.** Typically, the measure of central tendency used with grades is the mean.

**d.** You could use either the median or the mode. Using the mean is probably not appropriate because the numbers are somewhat arbitrary. *(Is excellent really three times better than poor?)*

**e.** Because the data are expressed in words rather than in numbers, you must use the mode.

### PRACTICE 2

Choose an appropriate measure of central tendency for the following data sets.

**a.** The prices of textbooks in your bookstore

**b.** The kinds of pets your classmates have

**c.** The number of credits students at your college take

### Answer Box

**Warm-up:**  1. \( s = 93 \)  2. \( t = 2 \)  **Practice 1:**  a. Mean is 0.74 million barrels per day, median is 0.5 million barrels per day, and mode is 0.5 million barrels per day.  b. The mean is larger than the median because the top oil-consuming countries are all outliers and lie far above the other countries.  **Practice 2:**  a. The mean or median would be appropriate.  b. Because the data are not numerical, the mode would be most appropriate.  c. Any of the three measures would be informative.
Reading Questions
1. The ____________ is the name given to the number that occurs most often in a data set.
2. The mean can be highly affected by ____________ in the data.
3. It is necessary to arrange the data in ascending or descending order in order to find the ____________.
4. The median of a data set with an even number of data points can be calculated by finding the ____________ of the middle two numbers.

ACTIVITY
Pulse Rates. The Activity in Section 10.1 presented a set of data consisting of the number of keys each individual in a class was carrying on the day of the survey. On another day, one of the authors surveyed his statistics class by asking them to calculate their pulse rates. They were to count their pulse for 10 seconds and then multiply the result by 6. Following are the results of that survey.

<table>
<thead>
<tr>
<th>60</th>
<th>66</th>
<th>90</th>
<th>84</th>
<th>78</th>
<th>8</th>
<th>60</th>
<th>60</th>
<th>72</th>
<th>66</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>60</td>
<td>72</td>
<td>84</td>
<td>72</td>
<td>72</td>
<td>54</td>
<td>72</td>
<td>84</td>
<td>66</td>
</tr>
<tr>
<td>72</td>
<td>96</td>
<td>72</td>
<td>72</td>
<td>60</td>
<td>54</td>
<td>156</td>
<td>72</td>
<td>60</td>
<td>72</td>
</tr>
</tbody>
</table>

1. Two of these items of data are obviously errors. Which ones are errors, and what do you think happened?
2. Enter the data in your calculator and sort in ascending order.
3. Calculate the mean, median, and mode of the data.
4. Now discard the two data items that seem incorrect and calculate the mean, median, and mode once again. Which of these measures of central tendency changed and why? Which remained the same?

10.2 Exercises
In Exercises 1 to 4, calculate the mean, median, and mode. Make an observation about the data.

1. Quiz scores
   a. 80, 80, 80, 85, 90, 95
   b. 96, 90, 85, 80, 80, 79
   c. 85, 82, 80, 65, 0, 0
   d. 0, 80, 85, 85, 90, 95
   e. What would you say about each student’s progress?

2. Advertised puppy prices, in dollars
   c. Golden retriever male: 750, 450, 450, 700, 350, 400, 400, 700
   d. Golden retriever female: 750, 525, 700, 400, 450, 450, 700

* Circled numbers are core exercises.
3. Advertised rent, in dollars, for one-bedroom apartments
   a. In a city with population 150,000: 375, 600, 475, 475, 475, 435, 435, 395, 545, 375, 375, 475, 415, 485, 525
   b. In a city with population 530,000: 535, 525, 525, 500, 550, 525, 525, 625, 525
4. Asking prices of used sport utility vehicles
   b. Three-year-old Ford Expedition: $19,975, $25,998, $22,988, $20,995, $23,000

● MEAN
5. Explain how to find the mean of a set of data.
6. Choose one: The mean is influenced by [all, most, few] numbers in the set. Explain.
7. If the mean is the same for two sets of data, does this imply that the numbers in the sets are the same? Explain.
8. Comment, using complete sentences, on the effect of one low grade or one high grade on the mean of the test scores.
9. How can the formula for the area of a trapezoid be rewritten to show a mean average? Explain your formula in words.

● MEDIAN
10. Explain how to find the median from a set of data.
11. Choose one: The median [is, is not] influenced by one large or small measurement. Explain.
12. Is it possible to have \( \frac{3}{4} \) of the students in a class score above the median test score? Explain.
13. Why might the median of a set of numbers and the median of a triangle have such similar names?
14. The guard rail or strip of ground between the traffic lanes of a freeway is called the median. Explain why this is an appropriate word.

● MODE
15. Explain how to find the mode from a set of data.
16. Mode has the same root as modern, model, and à la mode, which are associated with current, fashionable, or most popular styles. Explain how mode as an average fits with these other words.

● MEAN, MEDIAN, AND MODE
17. How does one large data value affect the mean? the mode? the median?
18. If a fly ball or strikeout is worth 0 and a hit is worth 1, is a baseball batting average (hits divided by times at bat) a median, mode, or mean?
19. When might the median provide a better description of the average than the mean?
20. When might the mean provide a better description of the average than the median?
21. What can you conclude about a set of data if the mean is larger than the median?
22. What can you conclude about a set of data if the median is larger than the mean?
23. List a set of numbers for which the mean is larger than the median. List a set of numbers for which the mean is smaller than the median. Explain how you created your list.
24. When does the median of a set of data lie below the mean? When does the median lie above the mean?
25. Ordinal data are ranked in terms of which value is higher or lower, but the ranking can be somewhat subjective. Ordinal rankings don’t allow us to state with certainty how much higher or lower one data value is than another. Give five examples of ordinal data and explain what you think would be the best measure of their center.
26. In Practice 1, the mean and median for oil usage are not very close together, so neither the median nor the mean gives a good summary of the data. If you wanted to use the mean in a summary description of the data, what else could you include in your description to make the summary more accurate?
27. The accompanying table includes the data from Table 5, as well as data for 2002, 2006, and 2010.

<table>
<thead>
<tr>
<th>Driver’s Age (years)</th>
<th>States in 1991</th>
<th>States in 2002</th>
<th>States in 2006</th>
<th>States in 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>20</td>
<td>11</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>16.5*</td>
<td>1</td>
<td>10</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>17</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>17.5*</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*Also includes fractional parts of a year other than \( \frac{1}{2} \).


c. Find the median and mean for each of the years.

28. A baseball batting average is the ratio of hits to times at bat:
   \[ \text{Batting average} = \frac{\text{number of hits}}{\text{number of times at bat}} \]

   Find the batting average, in decimal notation, for each of the following players. Round to the nearest thousandth.
   
   a. Roberto Clemente: 209 hits in 585 times at bat
   b. Juan Gonzalez: 152 hits in 584 times at bat (The hits included 43 home runs.)
   c. Ken Griffey, Jr.: 174 hits in 565 times at bat
   d. Ivan Calderon: 141 hits in 470 times at bat
   e. Hugh Duffy (1894): 236 hits in 539 times at bat
   f. Rogers Hornsby (1924): 227 hits in 536 times at bat

   Exercises 29 and 30 provide practice in calculating with weighted numbers.

29. Three colleges are competing in a track and field meet. A first-place finish is worth 5 points; second place, 3 points; third place, 1 point. GRCC has 4 first places, 3 seconds, and 2 thirds. LCC has 5 first places and 4 thirds. BCC has 1 first place, 7 seconds, and 4 thirds. What is each team’s total score?

30. A basketball team has 10 free throws at 1 point each, 25 field goals at 2 points each, and 6 three-point shots. The opposing team has 16 free throws, 20 field goals, and 4 three-point shots. What are the total scores?

The students in Exercises 35 to 38 are in the same course as those in Exercises 31 to 34. What final exam score does each student need in order to earn a 0.80 weighted average? A perfect final exam score is 1.00.

35. Tests: 0.76, 0.81, 0.72
36. Tests: 0.60, 0.65, 0.70
37. Tests: 0.50, 0.70, 0.70
38. Tests: 0.60, 0.70, 0.75

39. Project: Bookstore Discount. A local bookstore offers “Buy 12 books, get one free.” The store records each purchase on a card. After you buy 12 books, the store averages the purchases and you get to spend the average amount on your next book.

   a. What is the average value for a card that records purchases of 55, 6, 15, 40, 8, 35, 42, 6, 20, 15, 10, and 15? A card that records purchases of 10, 18, 40, 34, 12, 8, 50, 30, 45, 20, 10, and 5?
   b. What is the total value of the two free books from part a?
   c. Suppose you were able to sort the purchases and place the 12 largest on one card and the others on a second card. Could you increase the total dollar value of the two free books?
   d. Describe a way, if any, to sort the purchases to your advantage.

40. Project: Center of Population. The U.S. Center of Population is like a mean of where people live. It would be the balance point for the United States if each person (assumed to be of equal weight) were standing at her or his home location on a rigid plate the size of the country.

   a. Trace a small map of the United States, and guess where the population center started and how it moved during the period from 1790 to 2000. Explain your guesses.
   b. The center is recalculated after each ten-year census. Go to www.census.gov/geo/www/cenpop/meanctr.pdf and print the map. Explain changes in direction in the center.
Variation and Summary

OBJECTIVES

- Calculate measures of variation from sets of data.
- Draw and apply graphical displays of variation.

WARM-UP

Exercises 1 and 2 repeat the lists of the sale prices of homes in Newark, New Jersey, from the chapter opener. Calculate the mean and median of these prices in dollars.

1. Zip code 07102:
   110,000  150,000  150,000  157,000  172,000  180,000  195,000
   234,000  260,000  270,000  277,000  278,000  287,000  288,000
   290,000  305,000  350,000  375,000  430,000  432,000

2. Zip code 07104:
   300,000  330,000  212,000  260,000  63,000  275,000  132,000
   185,000  136,000  210,000  310,000  95,000  32,000  152,000
   200,000  350,000  41,000  297,000  237,000  250,000

Calculate the distance between these points.

3. (3, 7) and (−2, −5)
4. (−3, 5) and (5, −10)

Two patients’ heart rates are being monitored. One patient’s morning, noon, and evening rates were 71, 72, and 73, respectively. The other patient’s rates were 120, 72, and 24. Both have the same mean and median heart rates. Should they receive the same treatment?

In order to provide a better summary of the data, we need to include some measure of how the data are spread out. Clearly, the patients’ conditions are not adequately described by the mean and median (two measures of central tendency). In this section, we will examine measures of variation or dispersion, measures that describe how close to the middle a data value is or how scattered a set of data is.

MEASURES OF VARIATION OR DISPERSION

The range, five-number summary, and standard deviation are all tools that can help describe variation and thus improve a summary of the data.

RANGE

Finding the range is a first step in determining how sets of numbers differ. The range is calculated by subtracting the lowest number in the set from the highest number.

Range

The range is the difference between the largest and the smallest numbers in a data set:

\[ \text{Range} = \text{maximum number} - \text{minimum number} \]
EXAMPLE 1  Finding a range  Find the range for each zip code of the home sale prices listed in the Warm-up.

Solution  
Zip code 07102: $432,000 − $110,000 = $322,000  
Zip code 07104: $350,000 − $32,000 = $318,000

PRACTICE 1  What is the range of the patients’ heart rates?

FIVE-NUMBER SUMMARY  
The five-number summary is a list of the minimum, the lower quartile, the median, the upper quartile, and the maximum of a set of data. We will look at these one at a time.

Quartiles  One way to discuss variation in data is to divide the data into equal parts. If we divide a data set into four parts, with each part containing an equal number of observations (or as close to an equal number as possible), we are said to be finding quartiles.

The lower quartile, $Q_1$, is the median of the lower half of the data. The upper quartile, $Q_3$, is the median of the upper half of the data. The median of the entire data set is sometimes referred to as $Q_2$. If we divide the observations into 100 parts, we are finding percentiles. Thus, $Q_1$, $Q_2$, and $Q_3$ are the same as the 25th percentile, the 50th percentile, and the 75th percentile. If the number of observations is odd, we discard the middle observation when calculating the upper quartile and the lower quartile.

EXAMPLE 2  Finding quartiles  For the data in the Warm-up on New Jersey home sale prices for zip code 07102, find the following:

a. Median  
b. Lower quartile, $Q_1$  
c. Upper quartile, $Q_3$

Solution  

a. In the Warm-up in Section 10.1, we arranged the data in ascending order:

{110,000, 150,000, 150,000, 157,000, 172,000, 180,000, 195,000, 234,000, 260,000, 270,000, 277,000, 278,000, 287,000, 288,000, 290,000, 305,000, 350,000, 375,000, 430,000, 432,000}

In this set, there is no single middle number, so we average the 10th and 11th numbers, $270,000$ and $277,000$. The median is $273,500$.

b. $Q_1$ is $176,000$, the average of $172,000$ and $180,000$, the middle two of the first ten numbers.

c. $Q_3$ is $297,500$, the average of $290,000$ and $305,000$, the middle two of the last ten numbers.

PRACTICE 2  For the data in the Warm-up on New Jersey home sale prices in zip code 07104, find the following. (Don’t forget to arrange the data in ascending order.)

a. Median  
b. Lower quartile, $Q_1$  
c. Upper quartile, $Q_3$

Inter-quartile Range  The inter-quartile range (IQR) is the range from the first quartile to the third quartile: $IQR = Q_3 - Q_1$. The IQR is the width of the interval that contains the middle 50% of the data. The smaller the range, the less variation there is in the data. The inter-quartile range for the data in Example 2 is $297,500 − 176,000 = 121,500$. 
Minimum and Maximum  While the upper and lower quartiles give good information about the middle 50% of the data, they do not include any extreme values. To include some information about outliers in a descriptive summary, we also report the minimum and maximum numbers.

EXAMPLE 3  Summarizing with five numbers  Write the five-number summary for the home sale data for zip code 07102.

Solution  The minimum is $110,000; \( Q_1 \) is $176,000; the median is $273,500; \( Q_3 \) is $297,500; and the maximum is $432,000.

DISPLAYING DATA

Box and Whisker Plots  A box and whisker plot is a visual depiction of the five-number summary.

Drawing a Box and Whisker Plot

1. Locate the median on a horizontal line with a scale appropriate to the data set.
2. Draw a box (rectangle) from \( Q_1 \) to \( Q_3 \) on the horizontal line.
3. Draw a line from the left end of the box to the minimum number. Draw another line from the right end of the box to the maximum number. These lines form the whiskers.

EXAMPLE 4  Drawing a box and whisker plot  Find an appropriate scale and draw a box and whisker plot for the home sale data for zip code 07102.

Solution  Step 1: The minimum is $110,000. The maximum is $432,000. The horizontal line needs to cover a range of about $322,000. As with creating scales on axes, dividing by 10 and rounding to a convenient number is useful. We divide the horizontal line into units of $50,000 and begin with $100,000 as the leftmost starting point (see Figure 13). We write the numbers lightly in pencil (so that we can erase them later) and in thousands of dollars (so that we can avoid writing all the zeros). We then draw a vertical line at 273.5 (to indicate the median, $273,500).

\[ \text{Median on scaled line.} \]

Step 2: Next, we draw vertical lines at \( Q_1 = 176 \) and \( Q_3 = 297.5 \). We draw a rectangle between \( Q_1 \) and \( Q_3 \), as shown in Figure 14.

\[ \text{Box.} \]
Step 3: Finally, we mark the minimum, 110, and the maximum, 432, and connect them to the box as whiskers. We then erase any extra numbers. The final plot is shown in Figure 15.

![Figure 15](image_url)

**FIGURE 15** Box and whisker plot for house sales in zip code 07102.

In Exercise 25, you will draw a box and whisker plot for house sale data for zip code 07104 and compare it with Figure 15.

Sometimes box and whisker plots don’t display enough detail about how the data in a set vary. Bar graphs and histograms are other graphical displays that give a good idea of the variation of data.

**Bar Graphs** A bar graph, also called a bar chart, can be used to show the distribution of data when categories are not numerical. The bar graph in Figure 16 displays 2009 data on individuals’ reasons for moving, using horizontal rectangles of the same width. Each rectangle represents a single category. The length of each rectangle is proportional to the number of individuals in that category. Bars can also be vertical, a common choice when data are divided into a large number of categories.

![Figure 16](image_url)

**FIGURE 16** Movers by reason for moving in 2009 (excludes Armed Services). Family-related reasons for moving include a change in marital status or establishing one’s own household; work-related reasons include taking a new job or job transfer, looking for work, moving closer to work, or seeking an easier commute; housing-related reasons include wanting better or cheaper housing; other reasons include moving to attend college, for a change of climate, for health reasons, or as the result of a natural disaster.


**Histogram** A histogram is similar to a bar chart, but without gaps between rectangles. The data are divided into intervals (usually of equal size), and the area of a rectangle is proportional to the number of data values that fall into that interval. Histograms are used for numerical data that are continuous, such as measurements of height, weight, or time. Figure 17 shows the times for the top 100 female finishers in the 2010 New York City marathon.
The histogram shows how the finishing times vary. For example, 10 racers finished with times close to that of the winner, but not until after 180 minutes did so many people finish in one two-minute period.

### STANDARD DEVIATION

Box plots, bar charts, and histograms show the spread of data visually. A disadvantage of these summaries is that they do not give a numerical description of the data. One way to find a numerical measure of the spread of data is to consider the differences between each data point, \( x_i \), and the mean, \( \bar{x} \). The difference between each number and the mean is called the deviation.

In developing the formula for what British mathematician Karl Pearson was to call “standard deviation” in 1894, statisticians may have thought about how we find the distance between two points. The distance between \((x_1, y_1)\) and \((x_2, y_2)\) is given by

\[
d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}
\]

The distance formula finds the deviation between the \( x \)-coordinates and between the \( y \)-coordinates. To assure that the distances are positive, we square the deviations and then add them. To compensate for squaring the deviations, we then take the square root of the sum to obtain the distance. Observe the similarity between the distance formula and the formula for the sample standard deviation, a measure of the variation between the mean of a set and each number in the set.

#### Sample Standard Deviation

For a set of \( n \) numbers \( x_1, x_2, x_3, \ldots, x_n \) with mean \( \bar{x} \), drawn randomly from a population, the sample standard deviation is

\[
s_s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}
\]

\[
= \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}
\]
The division by \( n - 1 \) in the sample standard deviation formula creates an “average” deviation.

**EXAMPLE 5**

**Finding the standard deviation**

Find the standard deviation, to the nearest dollar, of the set of home sale price data for zip code 07102.

{ 110,000 150,000 150,000 157,000 172,000 180,000 195,000 234,000 260,000 270,000 277,000 278,000 287,000 288,000 290,000 305,000 350,000 375,000 430,000 432,000 } 

The mean is $259,500.

**Solution**

Finding the standard deviation requires subtracting the mean of the sample, $259,500, from each piece of data and then squaring the result. Only three of the subtracted and squared terms are shown under the radical sign below. The set of dots, "\( \ldots \)\), represents all the other terms.

\[
\sigma_x = \sqrt{\frac{(110,000 - 259,500)^2 + (150,000 - 259,500)^2 + \ldots + (430,000 - 259,500)^2}{20 - 1}} 
\]

\[
= 92,141
\]

The mean, sample standard deviation, and five-number summary are one-variable statistics. Measures that come in twos, such as ordered pairs \((x, y)\), are two-variable statistics. We will have more to say about these in Section 10.4.

**Graphing Calculator Technique: One-Variable Statistics**

1. Enter the data for home sales, \([\text{STAT}] 1 : \text{Edit}\), as shown in Figure 18. Place the data for zip code 07102 into L1 and the data for zip code 07104 into L2. (See Example 1 in Section 10.2, page 552.) Sort the lists separately into ascending order with \(2 : \text{SortA(}L1\text{)}\) and \(2 : \text{SortA(}L2\text{)}\). Sorting isn’t required, but it lets you scroll down the lists, as shown in Figure 19, and estimate \(Q_1\), the median, and \(Q_3\).

2. To calculate the one-variable statistics on L1 from the \([\text{STAT}] \text{CALC} 1 : 1–\text{Var Stats} \) menu (Figure 20), choose \([\text{STAT}] \text{CALC} 1 : 1–\text{Var Stats} 2\text{nd} \) \([L1]\) \([\text{ENTER}]\). The mean, \( \bar{x} \), will appear first in the list. The sample standard deviation, \( s_x \), is fourth in the list. Move the cursor down the screen for the
information below the arrow: minimum \( X \), quartiles \( (Q_1 \text{ and } Q_3) \), median, and maximum \( X \). (See Figures 21 and 22.)

\[
\begin{array}{|c|}
\hline
1\text{-Var Stats} \\
\bar{x}=259000 \\
\sum x=5190000 \\
\sum x^2=1.50811\times10^{12} \\
5x=92140.9104 \\
\sigma x=89807.85044 \\
\n\hline
\end{array}
\]

\[
\begin{array}{|c|}
\hline
1\text{-Var Stats} \\
\bar{x}=203350 \\
\sum x=4967000 \\
\sum x^2=1.0056\times10^{12} \\
5x=96945.6833 \\
\sigma x=94490.8858 \\
\n\hline
\end{array}
\]

**FIGURE 21** L1 statistics, top of screen.  **FIGURE 22** L1 statistics, bottom of screen.

Compare the results on the calculator with the results in Example 3.

3. To calculate the one-variable statistics on L2, repeat step 2 using \([L2]\) instead of \([L1]\). The results are shown in Figures 23 and 24.

\[
\begin{array}{|c|}
\hline
1\text{-Var Stats} \\
\bar{x}=203350 \\
\sum x=4967000 \\
\sum x^2=1.0056\times10^{12} \\
5x=96945.6833 \\
\sigma x=94490.8858 \\
\n\hline
\end{array}
\]

\[
\begin{array}{|c|}
\hline
1\text{-Var Stats} \\
\bar{x}=320000 \\
\sum x=3200000 \\
\sum x^2=1.34000 \times10^{12} \\
5x=211000 \\
\sigma x=286000 \\
\n\hline
\end{array}
\]

**FIGURE 23** L2 statistics, top of screen.  **FIGURE 24** L2 statistics, bottom of screen.

**PRACTICE 3**

Use Figures 23 and 24 to write the mean, standard deviation, and five-number sales summary for the data for zip code 07104.

**Answer Box**

**Warm-up:** 1. Mean, $259,500; median, $273,500
2. Mean, $203,350; median, $211,000
3. 13
4. 17
**Practice 1:** 2, 96  **Practice 2:** a. $211,000 b. $134,000 c. $286,000  **Practice 3:** Mean, $203,350; \sigma x$, $96,946; minimum, $32,000; Q_1, $134,000; median, $211,000; Q_3, $286,000; maximum, $350,000

**Reading Questions**

1. The standard deviation, range, IQR, and five-number summary all give information about the _________ of the data.

2. The IQR is the width of the interval containing the middle _________ % of the data.

3. What are the numbers in the five-number summary?

4. The calculation of the standard deviation uses the distance between a data point and the _________.

5. If we divide a data set into 100 equal parts, we have found the _________.
**ACTIVITY**

**Presidential Ages.** Is there any pattern in the ages of those elected to the U.S. presidency? The presidents of the United States are listed in Table 7 in chronological order, followed by their age at inauguration.

**TABLE 7** Presidential Age at Inauguration

<table>
<thead>
<tr>
<th>President</th>
<th>Age</th>
<th>President</th>
<th>Age</th>
<th>President</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Washington</td>
<td>57</td>
<td>Lincoln</td>
<td>52</td>
<td>Hoover</td>
<td>54</td>
</tr>
<tr>
<td>Adams, J.</td>
<td>61</td>
<td>Johnson, A.</td>
<td>56</td>
<td>Roosevelt, F.</td>
<td>51</td>
</tr>
<tr>
<td>Jefferson</td>
<td>57</td>
<td>Grant</td>
<td>46</td>
<td>Truman</td>
<td>60</td>
</tr>
<tr>
<td>Madison</td>
<td>57</td>
<td>Hayes</td>
<td>54</td>
<td>Eisenhower</td>
<td>62</td>
</tr>
<tr>
<td>Monroe</td>
<td>58</td>
<td>Garfield</td>
<td>49</td>
<td>Kennedy</td>
<td>43</td>
</tr>
<tr>
<td>Adams, J. Q.</td>
<td>57</td>
<td>Arthur</td>
<td>51</td>
<td>Johnson, L. B.</td>
<td>55</td>
</tr>
<tr>
<td>Jackson</td>
<td>61</td>
<td>Cleveland</td>
<td>47</td>
<td>Nixon</td>
<td>56</td>
</tr>
<tr>
<td>Van Buren</td>
<td>54</td>
<td>Harrison, B.</td>
<td>55</td>
<td>Ford</td>
<td>61</td>
</tr>
<tr>
<td>Harrison, W. H.</td>
<td>68</td>
<td>Cleveland</td>
<td>55</td>
<td>Carter</td>
<td>52</td>
</tr>
<tr>
<td>Tyler</td>
<td>51</td>
<td>McKinley</td>
<td>54</td>
<td>Reagan</td>
<td>69</td>
</tr>
<tr>
<td>Polk</td>
<td>49</td>
<td>Roosevelt, T.</td>
<td>42</td>
<td>Bush, G. H. W.</td>
<td>64</td>
</tr>
<tr>
<td>Taylor</td>
<td>64</td>
<td>Taft</td>
<td>51</td>
<td>Clinton</td>
<td>46</td>
</tr>
<tr>
<td>Fillmore</td>
<td>50</td>
<td>Wilson</td>
<td>56</td>
<td>Bush, G.</td>
<td>54</td>
</tr>
<tr>
<td>Pierce</td>
<td>48</td>
<td>Harding</td>
<td>55</td>
<td>Obama</td>
<td>47</td>
</tr>
<tr>
<td>Buchanan</td>
<td>65</td>
<td>Coolidge</td>
<td>51</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Enter the age data into your calculator in L1 and sort the data from lowest to highest.
2. Find the mean, median, and mode of the ages.
3. Calculate the standard deviation, IQR, and five-number summary for the ages.
4. Draw a box and whisker diagram for the age data.
5. One way statisticians define an outlier is as a data point that is more than 1.5 times the IQR above \( Q_3 \) or more than 1.5 times the IQR below \( Q_1 \). Write an expression for finding outliers and solve it for this data set. Which ages are outliers?
6. What restrictions does the Constitution place on the age of the president? How many IQRs is this value from the median?
7. After the vice president, next in line for president is the Speaker of the House. What restrictions does the Constitution place on the age of people elected to the House of Representatives? How many IQRs is this value from the median?

**Exercises**

1. Calculate the mean, the median, and the range of each set of incomes in Exercises 1 to 4.
   1. $8000, $10,000, $12,000, $13,000, $13,000, $100,000
   2. $4000, $8000, $9000, $9000, $100,000
   3. $20,000, $27,500, $27,500, $27,500, $27,500
   4. $23,000, $25,000, $26,000, $27,000, $29,000

   • Circled numbers are core exercises.

In Exercises 5 to 8, find \( Q_1 \) and \( Q_3 \) for each data set and draw a box and whisker plot.
5. Data from Exercise 1
6. Data from Exercise 2
7. Data from Exercise 3
8. Data from Exercise 4

In Exercises 9 to 12, find the sample standard deviation, \( s \). Round to the nearest hundred. Which data set would you expect to have the smallest \( s \)?
9. Data from Exercise 1
10. Data from Exercise 2
11. Data from Exercise 3
12. Data from Exercise 4
13. A building code requires that the risers of stairs in homes have no more than a $\frac{1}{8}$-inch variation in a flight of stairs. Suppose a flight of stairs has a riser-to-tread ratio of 7.5 in. to 10.5 in.
   a. What is the largest riser if the smallest is 7.5 inches?
   b. What is the smallest riser if the largest is 7.5 inches?
   c. Will the following set of steps pass inspection?
      \[
      \frac{7}{2}, \frac{7}{3}, \frac{7}{4}, \frac{7}{5}, \frac{7}{6}
      \]
      Explain in terms of the concepts of this section.
   d. List a set of five steps with an average 7.5-inch rise and a $\frac{1}{4}$-inch variation.
   e. List a set of five steps with an average 7.5-inch rise and a $\frac{1}{8}$-inch variation.
14. The average sale price of a home in one particular month in a particular city was $146,843. This average was calculated from sales of 325 homes. The median home price was $129,400.
   a. Why might the average be larger than the median?
   b. What was the total value of homes sold?
   c. Are we able to calculate the standard deviation from the given information?
   d. How many homes sold for less than $129,400?

In Exercises 15 to 18, find the mean, standard deviation, and five-number summary (Min, Q1, Med, Q3, Max) for the given sets of average monthly temperatures in the pairs of cities. Compare the cities.
15. a. Atlanta, GA: 43, 47, 54, 62, 70, 77, 80, 79, 73, 63, 53, 45
   b. Seattle, WA: 41, 43, 46, 50, 56, 61, 65, 66, 61, 53, 45, 41
16. a. Chicago, IL: 22, 27, 37, 48, 59, 68, 73, 72, 64, 52, 39, 27
   b. Hartford, CT: 26, 29, 38, 49, 60, 69, 74, 72, 63, 52, 42, 31
17. a. Honolulu, HI: 73, 73, 74, 76, 77, 80, 81, 82, 82, 80, 78, 75
   b. Miami, FL: 68, 69, 72, 76, 80, 82, 84, 84, 82, 79, 74, 70
18. a. Portland, OR: 40, 43, 47, 51, 57, 63, 68, 69, 64, 54, 46, 40
   b. Philadelphia, PA: 32, 35, 43, 53, 64, 72, 78, 76, 69, 57, 47, 37
19. Following are scores on a 100-point exam:
    78 85 67 89 93 88 76 73
    64 55 72 76 43 61 34 80
   a. Find the mean, median, mode, and standard deviation.
   b. Create a histogram for the data using a column width of 10: 0–9, 10–19, . . . .
20. The following data show the distribution of the 140 moons in our solar system. Find the quartiles for the number of moons of the eight planets plus Pluto. Find the mean and standard deviation.

<table>
<thead>
<tr>
<th>Planet</th>
<th>Moons</th>
<th>Planet</th>
<th>Moons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mercury</td>
<td>0</td>
<td>Saturn</td>
<td>33</td>
</tr>
<tr>
<td>Venus</td>
<td>0</td>
<td>Uranus</td>
<td>27</td>
</tr>
<tr>
<td>Earth</td>
<td>1</td>
<td>Neptune</td>
<td>14</td>
</tr>
<tr>
<td>Mars</td>
<td>2</td>
<td>Pluto</td>
<td>1</td>
</tr>
<tr>
<td>Jupiter</td>
<td>62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
21. The following table categorizes the scores earned by 30 students on a ten-point quiz.

<table>
<thead>
<tr>
<th>Score</th>
<th>Number of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

   Find the mean, median, mode, five-number summary, and standard deviation, and draw a box plot for these data.
22. The following data are scores on an IQ test.
    95 87 100 115 120 135 109 149 86 79
    100 103 98 114 125 130 148 76 92 107
   a. Create a frequency table for these data using the categories 70–79, 80–89, 90–99, 100–109, 110–119, 120–129, 130–139, and 140–149.
   b. Draw and carefully label a histogram for the data.
23. Compare the box plots and histograms of the following data sets.
    [20, 20, 20, 40, 40, 40, 60, 60, 60, 80, 80, 80]
    [20, 30, 40, 50, 60, 70, 80]
24. Listed below are air-quality index (AQI) levels of particulate matter (PM$_{2.5}$) and ozone for three cities, measured on the first day of each month from August 2010 to January 2011 (from www.airnow.gov). An AQI of 100 corresponds to the national air quality standard for the pollutant as set by the EPA. An AQI of 50 is healthy, whereas an AQI of 200 is not.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PM$_{2.5}$</td>
<td>85</td>
<td>67</td>
<td>62</td>
<td>47</td>
<td>28</td>
<td>52</td>
</tr>
<tr>
<td>Ozone</td>
<td>44</td>
<td>38</td>
<td>28</td>
<td>23</td>
<td>12</td>
<td>29</td>
</tr>
</tbody>
</table>

Central Los Angeles
Chapter 10

25. Using the same scale as in Figure 15 (page 564), make a box and whisker plot for house sale prices in zip code 07104. What do the two plots suggest about sales?

26. What is it about shoe sizes of men and women that makes the topic a popular classroom investigation worldwide? Who needs to know about the distribution of sizes? (Check classroom investigations for yourself: Search "box and whisker" along with "shoe size.")

27. Two box plots have the same numbers for $Q_1$ and for $Q_3$, but one plot has longer whiskers on each side. Compare the data.

28. How can you read the IQR from a box plot?

29. How would adding an outlier affect a box plot? adding a data point inside the IQR?

30. What is the median time of the top 100 runners in Figure 17?

31. Can we find the mean of a set of data from a box plot? from a histogram? (Use Figure 17 as an example.) Can we find the median of a data set from a box plot? from a histogram? If we can, explain how.

32. Can we tell if the mean is bigger or smaller than the median from a box plot? from a histogram?

33. What might have happened to home sales to cause the sharp rise in the red line in the following figure?

34. Project: Penny Plot

a. Plot the data given below, with date on the horizontal axis. (The first number is the date on the penny; the second is the weight of the penny in grams.)

<table>
<thead>
<tr>
<th>Year</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>1983D</td>
<td>2.501</td>
</tr>
<tr>
<td>1982D</td>
<td>2.518</td>
</tr>
<tr>
<td>1974</td>
<td>3.130</td>
</tr>
<tr>
<td>1968D</td>
<td>3.085</td>
</tr>
<tr>
<td>1977D</td>
<td>3.084</td>
</tr>
<tr>
<td>1985D</td>
<td>2.515</td>
</tr>
<tr>
<td>1984D</td>
<td>2.628</td>
</tr>
<tr>
<td>1973D</td>
<td>3.134</td>
</tr>
<tr>
<td>1991D</td>
<td>2.538</td>
</tr>
</tbody>
</table>

b. What do you observe from your graph? Find a way to use the mean and standard deviation to justify your observation.

35. Project: Alternative Box Plot. Because box plots are relatively new, most applications on the Web explain how they define their box plot. One variation on the box plot has whiskers stretching 1.5 times the distance \((Q_3 - Q_1)\) away from the box or to the maximum and minimum values, whichever is closer to the box. Data outside the whiskers are called outliers. The Australians who created the box plot below for their air pollution comparisons used this approach.

![Box Plot Example]


a. If \(Q_3 - Q_1 = 40\), what is the maximum length of a whisker in the alternative box plot?

b. Draw both types of box plot (the one described in the text and the one described above) for the data in Exercise 1. Name the outliers.

c. On the Internet, find an example of a box plot of data from another field.

### 10.4 Linear Regression: Finding Equations from Data

#### Objectives
- Make a scatter plot.
- Fit a line through several data points.
- Use a calculator to find a line of best fit.
- Use a regression line to make predictions and estimations.
- Recognize the difference between correlation and causation.

#### Warm-Up

1. Find the slope of the line through each of the following sets of points.
   a. \((10, 6.79)\) and \((25, 6.99)\)
   b. \((25, 6.99)\) and \((50, 9.89)\)
   c. \((50, 9.89)\) and \((100, 13.99)\)

2. Find an equation of the line through \((10, 6.79)\) with slope 0.08.

A graph of data involving two variables is called a scatter plot. If we graph data from applications on a scatter plot, the data seldom fall into perfect mathematical shapes: straight lines or known curves. However, when the scatter plot
suggests that data might be related by a line or a curve, we can find an equation that approximately fits the data. If it is the equation of a _line that approximately fits the data_, that line is called a _line of best fit_ or _regression line_.

**ESTIMATING AN EQUATION FROM A TABLE**

Items in hardware stores are often available in a variety of sizes, which relate to prices. Suppose a new department supervisor wants to examine current prices before setting new prices. Building an equation from existing prices is one strategy. In Example 1, we estimate an “average” slope from slopes between individual length and price pairs and use this slope to write an equation.

**EXAMPLE 1** Finding a linear equation with a table  A local hardware store has these prices for medium-duty extension cords: 10 feet, $6.79; 25 feet, $6.99; 50 feet, $9.89; 100 feet, $13.99.

a. Draw a scatter plot of the data.

b. Suggest whether a linear equation might describe the data.

c. Use a table to estimate the slope and find a y-intercept and write a linear equation.

d. What does the equation mean in the problem setting?

**Solution**

a. The scatter plot is shown in Figure 25.

![Figure 25](image)

**b.** Yes; the four points lie close to a line, even though the second point is a bit low.

c. Table 8 lists the slopes: 0.013, 0.116, and 0.082. An easy estimate of the overall slope is 0.08, slightly lower than the last slope value.

<table>
<thead>
<tr>
<th>Δx</th>
<th>Length, x (feet)</th>
<th>Price, y (dollars)</th>
<th>Δy</th>
<th>Slope, Δy/Δx</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td>6.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>25</td>
<td>6.99</td>
<td>0.2</td>
<td>0.2/15 ≈ 0.013</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>9.89</td>
<td>2.9</td>
<td>2.9/25 = 0.116</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>13.99</td>
<td>4.1</td>
<td>4.1/50 = 0.082</td>
</tr>
</tbody>
</table>

**FIGURE 25** Extension cord price.

---

89347_ch10_ptg01_ch01_AIE4.indd   572
12/9/11   9:19 AM
If we substitute \((x, y) = (10, 6.79)\) and \(m = 0.08\) into \(y = mx + b\), we can solve for \(b\).

\[
6.79 = (0.08)(10) + b \\
b = 5.99
\]

The resulting linear equation is \(y = 0.08x + 5.99\), shown in Figure 26.

**Figure 26**

- The equation suggests that all the extension cord prices include a common cost (end parts, assembly, packaging, and profit for the store) of $5.99 plus a cost of $0.08 per foot (materials and shipping).

In Example 1, we could have used any point to calculate the vertical intercept. For example, if we had chosen the point \((50, 9.89)\) instead of the point \((10, 6.79)\), we would have calculated the vertical intercept as follows:

\[
9.89 = 0.08(50) + b \\
9.89 = 4.0 + b \\
5.89 = b
\]

The resulting line, \(y = 0.08x + 5.89\), has the same slope but a different \(y\)-intercept, as shown in Figure 27.
PRACTICE 1

Repeat parts a to d in Example 1 for prices in 2010 of Pyrex glass measuring cups:
1 cup, $3.59; 1 pint, $4.49; 1 quart, $5.79; 2 quarts, $8.29.

Summary: Finding a Linear Equation from a Table

1. Graph the data to see whether they have a linear shape.
2. Use the change in $x$ and change in $y$ to find the slope between ordered pairs. Create a table of these values. Estimate an average slope, $m$, for the table.
3. Substitute the estimated slope and any point from the table into the slope-intercept form to find the vertical intercept.
4. State the equation in the form $y = mx + b$.
5. Graph the equation to make sure that it is reasonably close to the data.

FITTING AN EQUATION WITH LINEAR REGRESSION

The graphing calculator statistical functions include a feature called linear regression, or LinReg, to fit a regression line to data. Although you may be surprised at the accuracy with which you can find equations with a table, there is no question that for large amounts of data or for “messy” numbers, a calculator approach is welcome.

Graphing Calculator Technique: Linear Regression

1. Begin by pressing STAT ENTER.
2. Clear prior lists L1 and L2. To clear, move the cursor up to L1, press CLEAR, and then move the cursor down; the list will be emptied. Repeat as needed for other lists.
3. Enter the new data, usually in L1 and L2. Set up WINDOW and graph using 2ND [STAT PLOT].

Because LinReg uses $y = ax + b$ instead of $y = mx + b$, use $a$ in place of $m$.

EXAMPLE 2

Finding an equation with linear regression

Using linear regression, fit a linear equation to the extension cord price data from Example 1: 10 feet, $6.79; 25$ feet, $6.99; 50$ feet, $9.89; 100$ feet, $13.99$.

a. Enter the data into L1 and L2. Use the calculator to draw a scatter plot. Verify that a linear regression is appropriate.
b. Perform the linear regression. State the equation, rounding to three decimal places.
c. Graph the regression equation along with the data and graph of $y = 0.08x + 5.99$ from Example 1.

Solution

a. Press STAT ENTER on the TI-84 to obtain the data entry and edit screen. Figure 28 shows the extension cord data with the length of each cord in L1 and price of each cord in L2. To set dimensions of the graph, press WINDOW and enter intervals for $x$ and $y$. Include all lengths and prices within the window intervals, as in Figure 29.
Before obtaining the scatter plot, clear prior equations. Press \( \text{Y=} \), move the cursor to the first number or letter of each equation, and press \( \text{CLEAR} \). To obtain the scatter plot, press \( \text{2ND} \) [STAT PLOT], which will give Figure 30, and then press \( \text{ENTER} \) twice (first to turn Plot1 on and second to obtain options for Plot1). Change options, as needed, to match the screen in Figure 31.

To perform the linear regression, first press \( \text{STAT} \), choose \( \text{CALC} \), then choose 4, as shown in Figure 33. (If you move the cursor to 4 : LinReg, you will need to also press \( \text{ENTER} \).) LinReg(ax + b) will appear on the home screen, as in Figure 34. Press \( \text{2ND} \) [L1] \( \text{3} \) \( \text{2ND} \) [L2] for the list shown in Figure 35. Bonus: You can automatically have the regression equation placed in Y = by going to variable storage and choosing Y1 with \( \text{Y-VARS} \) \( \text{1} \) \( \text{ENTER} \). The regression equation is shown in Figure 36. Rounding to three decimal places, it is \( y = 0.084x + 5.508 \).

Your calculator may display \( r^2 \) and \( r \). This information is useful in higher level statistics courses.

To graph the equations, press \( \text{Y=} \). The regression equation is already in Y1, as shown in Figure 36. Enter the equation from Example 1: \( Y_2 = 0.08X + 5.99 \). Press \( \text{GRAPH} \). The graphs are shown in Figure 37. If you are unclear as to which is the regression line, press \( \text{TRACE} \). The up or down cursor will highlight the data and the line for each equation, as in Figure 38. The equation will be in the upper left corner.
Use a calculator to find a linear regression equation for the 2010 prices of Pyrex glass measuring cups: 1 cup, $3.59; 1 pint, $4.49; 1 quart, $5.79; 2 quarts, $8.29. State the window dimensions in which you view the scatter plot of the data and the graph of the regression equation.

**EXAMPLE 3** Estimating a number within a set of data  
In Example 2, we found the linear regression equation $y = 0.084x + 5.508$ for the variation of the price, $y$, with the number of feet, $x$, of extension cord. As store manager, you need to price a 75-foot extension cord.

**Solution**  
Because $x$ is the number of feet of extension cord, let $x = 75$. Then

$$y = 0.084(75) + 5.508 = 11.808$$

The price of a 75-foot cord is $11.81.

Because we used the regression equation to estimate a value within the set of data, the process is called interpolation. Estimates arrived at through interpolation are usually reliable.

However, suppose you wanted to price a 200-foot extension cord and you substituted $x = 200$ into the regression equation:

$$y = 0.084(200) + 5.508 = 22.308$$

The price of a 200-foot cord would be $22.31. Here we are estimating using an input outside the range of the data. This process is called extrapolation. Estimates arrived at through extrapolation are usually not reliable, especially if the input lies far away from the original data.

**EXAMPLE 4** Fitting linear equations and making predictions  
Table 9 shows the average sale price of all homes sold in Seattle, Washington, for each of the given years:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sale Price</td>
<td>230,000</td>
<td>246,000</td>
<td>305,000</td>
<td>355,000</td>
<td>410,000</td>
</tr>
</tbody>
</table>

It is often more convenient to use the number of years after 2000, rather than the actual year, as the independent variable. Table 10 incorporates this change.
### Section 10.4 Linear Regression: Finding Equations from Data

#### Table 10 Home Prices in Seattle, Years after 2000

<table>
<thead>
<tr>
<th>Years after 2000</th>
<th>0</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sale Price</td>
<td>230,000</td>
<td>246,000</td>
<td>305,000</td>
<td>355,000</td>
<td>410,000</td>
</tr>
</tbody>
</table>

**a.** Fit a linear regression line to the data in Table 10, with the average sale price of a home varying with the number of years after 2000.

**b.** Use the regression equation to estimate (interpolate) the average cost of a home in Seattle, Washington, in 2004.

**c.** Use the regression equation to predict (extrapolate) the average cost of a home in Seattle, Washington, in 2010.

**Solution**

**a.** $P = 23,093.14t + 221,446.08$, where $P$ is the average sale price of a home and $t$ is the years after 2000

**b.** Letting $t = 4$ in the regression equation, we obtain $313,819$. The actual average price in 2004 was $320,000$. Our interpolation is fairly accurate.

**c.** Letting $t = 10$ in the regression equation, we obtain $452,377$. The actual average price was $340,000 in 2010. The extrapolation is not reliable. When we use the linear regression equation to make predictions outside the data range, we are assuming that the trend established will continue. As you may know, there was a serious real estate market crash in late 2008 and prices fell nationwide.

#### Example 5 Applying a Linear Regression Equation

Consider the data in Table 11, which shows sales of music CDs in the United States for given years since 2000.

#### Table 11 Sales of Music CDs since 2000

<table>
<thead>
<tr>
<th>Years after 2000</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD Sales (in millions)</td>
<td>1013</td>
<td>882</td>
<td>803</td>
<td>746</td>
<td>720</td>
</tr>
</tbody>
</table>

**a.** Using a calculator, find a linear regression line that expresses how CD sales vary with the number of years after 2000.

**b.** Use this regression line to predict the number of CDs sold in the United States in 2014.

**c.** Now draw a scatter diagram of these data. Use X[−1, 5], Xscl 1; Y[600, 1100], Yscl 100.

**Solution**

**a.** $y = −72.2x + 977.2$

**b.** $−33.6$. This answer is obviously not valid. While it is possible that CD sales will continue to fall, they cannot be negative.

**c.** The data are shown in Figure 39.

![Figure 39: CD sales.](image)

A close look at this screen shows that a line will not fit the data very well. This suggests that you should look at a scatter plot of the data prior to calculating a linear regression line.
The examples have illustrated important points that can be summarized as follows.

### Linear Modeling with a Calculator

1. Create a scatter plot to see if the data are approximately linear.
2. Use a calculator to find the line of best fit.
3. Make sure the line of best fit is close to most of the data points.
4. Use the line of best fit to make predictions and estimations.
5. Be careful when extrapolating from the data set. These predictions may not be valid.

### CORRELATION

#### EXAMPLE 6

**Applying linear regression** The data in Table 12 compare the amount various communities spend per capita on police services with the crime rate (in number of offenses reported to police).

<table>
<thead>
<tr>
<th>Amount Spent on Police per Capita (dollars)</th>
<th>149</th>
<th>109</th>
<th>82</th>
<th>115</th>
<th>65</th>
<th>71</th>
<th>121</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime Rate (per million people)</td>
<td>196.9</td>
<td>123.4</td>
<td>96.3</td>
<td>155.5</td>
<td>85.6</td>
<td>70.5</td>
<td>167.4</td>
<td>84.9</td>
</tr>
</tbody>
</table>

*Source: FBI Uniform Crime Report.*

a. Draw a scatter plot for these data. Does the plot suggest that a linear equation will fit the data?

b. Find the linear regression line expressing how the crime rate varies with the amount spent on police.

c. Does this data set seem to you to make sense?

**Solution**

a. The scatter plot is shown in Figure 40.

![Figure 40](image)

The data seem to have a generally linear pattern.

b. \( R = 1.515s - 26.453 \), where \( R \) is the crime rate and \( s \) is the amount spent on police per capita.

c. We hope the data set doesn’t make sense! Doesn’t it seem odd, even ridiculous, that as expenditures on police increase, the crime rate also increases? There must be other factors that influence crime rate instead of expenditures on police. It may be coincidental that these data fit a line. This example points out the difference between linear correlation and causation.
Correlation

Correlation is a statistical description of the degree to which two (or more) variables vary together. Correlation generally lacks the precisely defined relationship implied by a function. A positive correlation indicates that two sets of numbers simultaneously increase, and a negative correlation indicates that when the numbers in one set increase, the numbers in the other set decrease. Linear correlation indicates that the data vary in a linear fashion. Most importantly, linear correlation does not, by itself, suggest causation (a cause-and-effect relationship). One way to determine whether a correlation is a causation is to conduct a controlled statistical study.

PRACTICE 3

Table 13 shows U.S. energy consumption from renewable resources, in quadrillion BTUs, for given years.

<table>
<thead>
<tr>
<th>Year</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy Consumed</td>
<td>4.56</td>
<td>4.67</td>
<td>5.35</td>
<td>6.21</td>
<td>5.4</td>
<td>6.15</td>
<td>6.261</td>
<td>6.44</td>
<td>6.92</td>
<td>6.83</td>
</tr>
</tbody>
</table>

a. Draw a scatter plot, using years after 1998 as the independent variable and energy consumption as the dependent variable.

b. Does it seem as though a linear equation would fit the data well?

c. Fit a linear equation to the data, using linear regression on a calculator.

d. Use the regression line to estimate energy consumption from renewable resources in 2012.

e. Predict the year in which energy consumption from renewable resources will be 10 quadrillion BTUs.

f. Predict the year in which energy consumption from renewable resources would have been 0. Does this seem reasonable?

Answer Box

Warm-up: 1. a. \( m = 0.013 \)  b. \( m = 0.116 \)  c. \( m = 0.082 \)

2. \( y = 0.08x + 5.99 \)

Practice 1: a. [Graph]

b. The points line up well.  c. The following table shows slopes 0.625, 0.65, and 0.9. The average slope is \( m = 0.7 \). Substituting \((4, 5.79)\) and \( m = 0.7 \) into \( y = mx + b \) and solving for \( b \) gives \( b = 2.99 \). One equation is \( y = 0.7x + 2.99 \). Other points give \( b = 2.89 \), \( b = 3.09 \), and \( b = 2.69 \), respectively.
(concluded)

<table>
<thead>
<tr>
<th>Δx</th>
<th>Volume, x (cups)</th>
<th>Price, y (dollars)</th>
<th>Δy</th>
<th>Slope Δy/Δx</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3.59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4.49</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>5.79</td>
<td>1.3</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>8.29</td>
<td>2.5</td>
<td>0.625</td>
</tr>
</tbody>
</table>

d. The slope is the price for each additional unit of volume (cup), and the y-intercept would be the price for a zero-capacity measuring cup, which would not be meaningful. **Practice 2**: \( y = 0.659x + 3.068 \).

**Window**: X[0, 10], scl 2, and Y[0, 10], scl 2

![Graph](image)

**Practice 3**: a.

b. While the data seem somewhat scattered, energy consumption did rise over time. **c.** \( y = 0.258x + 4.719 \) **d.** 8.33 quadrillion BTUs **e.** In about 20.5 years from 1998, or 2018 **f.** About 18.3 years prior to 1998, or 1980. This does not seem reasonable, because hydroelectric power is a renewable resource and the massive Hoover Dam was built in the 1930s, well before 1980.

**Reading Questions**

1. To estimate outside the range of data, you use ________: to estimate within a set of data, you use ________.
2. The first step in linear modeling is to create a (an) ________.
3. Linear correlation does not by itself suggest ________.
4. A graph of data involving two variables is a (an) ________.
5. ________ is the term for a line of best fit found by using a calculator.

**ACTIVITY**

**Airport Shuttle.** You own a shuttle service that transports passengers from their homes to the airport. You use seven-passenger vans, and the time has come to replace some of your fleet. You are considering buying Chrysler Town and Country vans. How well these vans retain...
their value and how long you can reasonably keep them are questions of interest. To help you answer these questions, you investigate the resale value in different years of a 2006 model of this van. You believe that by investigating these data you will be able to determine how well a current model will retain its value. The resale values of a 2006 model Chrysler Town and Country are given in Table 14.

**TABLE 14** Resale Values of Chrysler Van

<table>
<thead>
<tr>
<th>Year</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years after 2006</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Resale Value</td>
<td>26,340*</td>
<td>17,550</td>
<td>12,525</td>
<td>10,325</td>
<td>8950</td>
</tr>
</tbody>
</table>

*Original purchase price of new van in 2006.

1. Draw a scatter plot for the data.

2. Use a calculator to find a linear regression equation, with resale value varying with years after 2006.

3. What is the slope of the line? What does this slope represent in the problem setting?

4. Use your linear model to predict the resale value of the van in 2011.

5. Find the intercepts of your linear regression line. What do these intercepts represent in the problem?

6. The van is considered to be more expensive to maintain than it is worth when its resale value is $3500. Use your linear regression line to determine when this occurs.

7. Using the information from parts 1 to 6, make an argument about whether the line is a good fit for the data.

### Exercises

**10.4**

1. In 2006, the prices of Pyrex glass measuring cups were as follows: 1 cup, $3.19; 2 cups, $3.99; 4 cups, $4.99; 8 cups, $7.49. Compare the data here with those in Practice 1 as needed.
   a. Place the 2006 volume and price data in a table. Estimate an average slope, and use the slope-intercept form to find an equation.
   b. Compare the equations for 2010 and 2006 in terms of the prices.

2. In 2010, Heuck Classic mixing bowls were priced as follows (rounded to the nearest dollar): 1 quart at $3, 3 quarts at $5, 5 quarts at $8, and 8 quarts at $10. Place the volume and price data in a table. Estimate an average slope, and use the slope-intercept form to find an equation.

In Exercises 3 to 6, draw a scatter plot, find a linear regression equation if appropriate, predict the indicated costs, and suggest why the actual cost might differ from your calculation.


5. A can of Comet powdered cleanser cost $0.34 in 1986, $0.89 in 2003, and $0.99 in 2006. Predict the cost in 2010 and compare the prediction with the actual cost of $1.19.

7. Cost of Wood Dowels at a Hobby Store

<table>
<thead>
<tr>
<th>Diameter (inches)</th>
<th>1/8</th>
<th>1/4</th>
<th>1/2</th>
<th>3/4</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost (dollars)</td>
<td>0.20</td>
<td>0.25</td>
<td>0.69</td>
<td>1.47</td>
<td>2.19</td>
</tr>
</tbody>
</table>

a. Draw a scatter plot for diameter and cost. Is a linear regression appropriate?
b. Draw a scatter plot for cross-sectional area and cost. Is a linear regression appropriate?
c. Which is the better predictor of cost: diameter or area?
d. Predict the cost of a $\frac{5}{16}$-inch-diameter dowel.

8. Plastic Tarps at a Local Camping Store

<table>
<thead>
<tr>
<th>Dimensions (feet)</th>
<th>5 × 7</th>
<th>8 × 10</th>
<th>10 × 12</th>
<th>12 × 14</th>
<th>18 × 24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost (dollars)</td>
<td>3.99</td>
<td>7.99</td>
<td>11.97</td>
<td>16.97</td>
<td>39.97</td>
</tr>
</tbody>
</table>

a. Draw a scatter plot for width (the first of the two dimensions) and cost. Is a linear regression appropriate?
b. Draw a scatter plot for area and cost. Do the linear regression, if appropriate.
c. Which is the better predictor of cost: width or area?
d. Predict the cost of a 10 × 18 tarp. Store price was $17.97. Use the answer to part b.
e. Predict the cost of a 20 × 30 tarp. Store price was $49.97. Use the answer to part b.

9. A Macy’s newspaper insert offered a special on diamond stud earrings.

Diamond Studs in 14K White Gold

<table>
<thead>
<tr>
<th>L1: Weight (carats)</th>
<th>$\frac{1}{3}$</th>
<th>$\frac{3}{4}$</th>
<th>1</th>
<th>$1\frac{3}{4}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2: Regular Price (dollars)</td>
<td>500</td>
<td>1900</td>
<td>2700</td>
<td>3900</td>
</tr>
<tr>
<td>L3: Sale Price (dollars)</td>
<td>349</td>
<td>1329</td>
<td>1889</td>
<td>2729</td>
</tr>
</tbody>
</table>

a. Draw a scatter plot for weight and regular price, L1 and L2. Do the linear regression, if appropriate.
b. Draw a scatter plot for weight and sale price, L1 and L3. Do the linear regression, if appropriate.
c. Draw a scatter plot for regular price and sale price, L2 and L3. (This will require a new window.) Do the linear regression, if appropriate. Interpret the equation.

d. According to your linear regression line, when will the population of Casper, Wyoming, reach 55,000?

10. In spite of the availability of the Internet, books such as The World Almanac are still popular. Following are data on edition years and cover prices: 1995, $8.95; 2000, $10.95; 2003, $11.95; and 2006, $12.95

a. Draw a scatter plot for the time in years after 1995 and the cover price.
b. Fit a linear equation to the data, if appropriate.
c. Extrapolate to the 15th year, 2010, to predict a price.
d. The actual price for the 2010 edition was $12.99. Suggest reasons why the price was almost $1.50 lower than predicted.

11. Estimates by the U.S. Census Bureau of the population of Casper, Wyoming, are given in the table below.

<table>
<thead>
<tr>
<th>Year</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>49,644</td>
</tr>
<tr>
<td>2001</td>
<td>49,831</td>
</tr>
<tr>
<td>2002</td>
<td>50,124</td>
</tr>
<tr>
<td>2003</td>
<td>50,566</td>
</tr>
<tr>
<td>2004</td>
<td>50,944</td>
</tr>
<tr>
<td>2005</td>
<td>51,498</td>
</tr>
<tr>
<td>2006</td>
<td>51,965</td>
</tr>
<tr>
<td>2007</td>
<td>53,005</td>
</tr>
</tbody>
</table>

Source: quickfacts.census.gov.

a. Find the linear regression line that best fits these data, using years after 2000 as your input variable.
b. What do the slope and y-intercept of your line mean in the context of the problem?
c. Use your linear regression line to predict the population of Casper, Wyoming, in 2012.
d. According to your linear regression line, when will Casper, Wyoming’s population reach 55,000?

12. The U.S. Census Bureau has compiled estimates of the number of centenarians (individuals over 100 years of age) living in the United States in various years. The data are shown in the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Estimated Number of Centenarians</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>2300</td>
</tr>
<tr>
<td>1960</td>
<td>3300</td>
</tr>
<tr>
<td>1970</td>
<td>4800</td>
</tr>
<tr>
<td>1980</td>
<td>15,000</td>
</tr>
<tr>
<td>1990</td>
<td>32,000</td>
</tr>
<tr>
<td>2000</td>
<td>50,500</td>
</tr>
<tr>
<td>2010</td>
<td>70,490</td>
</tr>
</tbody>
</table>
a. Draw a scatter plot of these data. Is a linear regression appropriate?

b. Find the linear regression line that best fits the data. Use years after 1950 as your input variable.

c. Use your linear regression line to predict the number of U.S. centenarians at the next census, 2020.

d. According to your linear regression line, when will there be 100,000 centenarians in the United States?

13. The following data set relates the number of cigarettes smoked annually per capita (in hundreds) in ten randomly selected states to lung cancer deaths (per 100,000 people) in that state.

<table>
<thead>
<tr>
<th>State</th>
<th>Cigarette Smoking Rate (in hundreds per capita)</th>
<th>Lung Cancer Deaths (per 100,000 people)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>18.2</td>
<td>17.05</td>
</tr>
<tr>
<td>Arizona</td>
<td>25.82</td>
<td>19.8</td>
</tr>
<tr>
<td>Mississippi</td>
<td>15.08</td>
<td>16.03</td>
</tr>
<tr>
<td>Idaho</td>
<td>20.1</td>
<td>13.58</td>
</tr>
<tr>
<td>Indiana</td>
<td>26.18</td>
<td>20.3</td>
</tr>
<tr>
<td>Iowa</td>
<td>22.12</td>
<td>16.59</td>
</tr>
<tr>
<td>Kansas</td>
<td>21.84</td>
<td>16.84</td>
</tr>
<tr>
<td>Missouri</td>
<td>27.56</td>
<td>20.98</td>
</tr>
<tr>
<td>South Dakota</td>
<td>20.94</td>
<td>14.11</td>
</tr>
<tr>
<td>Alaska</td>
<td>30.34</td>
<td>25.88</td>
</tr>
</tbody>
</table>


a. Draw a scatter plot of the data. Does it look as if a linear regression might model the data accurately?

b. Find the linear regression line that best fits the data. Use weight as your input variable.

c. What do the slope and y-intercept of your line mean in the context of the problem?

d. Use your linear regression line to predict the miles per gallon of a Hyundai Santa Fe that weighs 3875 pounds. How closely does your prediction match the current rating of this car, which is 20 miles per gallon?

e. According to this model, how much would a car have to weigh for its mileage rating to be 30 mpg?

14. The following data set gives the weight (in thousands of pounds) and the city miles per gallon rating of various cars.

<table>
<thead>
<tr>
<th>Weight (thousands of pounds)</th>
<th>Miles per gallon</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.36</td>
<td>16.9</td>
</tr>
<tr>
<td>4.054</td>
<td>15.5</td>
</tr>
<tr>
<td>3.605</td>
<td>19.2</td>
</tr>
<tr>
<td>3.41</td>
<td>18.1</td>
</tr>
<tr>
<td>3.84</td>
<td>17</td>
</tr>
<tr>
<td>3.725</td>
<td>17.6</td>
</tr>
<tr>
<td>3.955</td>
<td>16.5</td>
</tr>
<tr>
<td>3.83</td>
<td>18.2</td>
</tr>
<tr>
<td>2.585</td>
<td>26.5</td>
</tr>
<tr>
<td>2.67</td>
<td>27.4</td>
</tr>
<tr>
<td>2.6</td>
<td>21.5</td>
</tr>
<tr>
<td>1.925</td>
<td>31.9</td>
</tr>
</tbody>
</table>

Source: Congressional Budget Office.

a. Draw a scatter plot of these data. Is a linear regression appropriate?

b. Find the linear regression line that best fits the data. Use years after 2000 as your input variable.
c. What do the slope and y-intercept of your line mean in the context of the problem?

d. Use your linear regression line to model the amount of public debt in 2012.

e. According to this model, when will public debt exceed $10,000 billion?

16. The data set below shows the timber harvest in Washington state from 1990 to 2006, in millions of board feet.

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Year</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>5849</td>
<td>1999</td>
<td>4383</td>
</tr>
<tr>
<td>1991</td>
<td>5104</td>
<td>2000</td>
<td>4177</td>
</tr>
<tr>
<td>1992</td>
<td>5018</td>
<td>2001</td>
<td>3716</td>
</tr>
<tr>
<td>1993</td>
<td>4329</td>
<td>2002</td>
<td>3582</td>
</tr>
<tr>
<td>1994</td>
<td>4086</td>
<td>2003</td>
<td>4234</td>
</tr>
<tr>
<td>1995</td>
<td>4392</td>
<td>2004</td>
<td>3946</td>
</tr>
<tr>
<td>1996</td>
<td>4249</td>
<td>2005</td>
<td>3730</td>
</tr>
<tr>
<td>1997</td>
<td>4245</td>
<td>2006</td>
<td>3483</td>
</tr>
<tr>
<td>1998</td>
<td>4022</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: www.dnr.wa.gov

a. Draw a scatter plot of these data. Is a linear regression appropriate?
b. Find the linear regression line that best fits the data. Use years after 1990 as your input variable.
c. What do the slope and y-intercept of your line mean in the context of the problem?
d. Use your linear regression line to predict the timber harvest in Washington in millions of board feet in 2010.
e. According to this model, when will the total timber harvest fall below 2500 million board feet?

17. Years ago, one of the authors asked five friends to divulge their shoe sizes and their annual salaries. After numbers were rounded to the nearest whole number for shoe size and nearest thousand dollars for salary, the results were (8, 25,000), (9, 36,000), (10, 45,000), (11, 52,000), and (12, 63,000).

a. Draw a scatter plot of these data. Does it look as though a linear equation will fit?
b. Find the linear regression line using a calculator.
c. Use this linear regression line to predict the annual salary of a person who wears a size 7 shoe. Do you believe this estimate?

18. The following data show per capita milk consumption and per capita carbonated soda consumption per year in the United States. The amounts are in gallons per person per year.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>27.5</td>
<td>26.6</td>
<td>25.7</td>
<td>23.9</td>
<td>22.5</td>
<td>22.0</td>
<td>21.9</td>
<td>21.6</td>
<td>27.5</td>
</tr>
<tr>
<td>Soda</td>
<td>35.1</td>
<td>35.7</td>
<td>46.2</td>
<td>47.4</td>
<td>49.3</td>
<td>46.7</td>
<td>46.6</td>
<td>46.4</td>
<td>35.1</td>
</tr>
</tbody>
</table>

Source: Statistical Abstract of the United States

a. Draw a scatter plot of the milk consumption and soda consumption data. Is a linear regression appropriate?
b. Find the linear regression line that best fits the data. Use milk consumption as your input variable.
c. What do the slope and y-intercept of your line mean in the context of the problem?
d. What is the estimate for soda consumption if milk consumption is 20 gallons per year?
e. Do you believe this is a correlation relationship or causation relationship? Explain.

19. The data below show per capita red meat consumption and per capita poultry consumption in the United States for selected years. The units are pounds per person per year.

<table>
<thead>
<tr>
<th>Year</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poultry</td>
<td>40.8</td>
<td>45.6</td>
<td>56.2</td>
<td>62.1</td>
<td>67.9</td>
<td>67.8</td>
<td>70.8</td>
<td>71.3</td>
<td>72.8</td>
</tr>
<tr>
<td>Red Meat</td>
<td>113.7</td>
<td>111.4</td>
<td>114.1</td>
<td>111.8</td>
<td>112.2</td>
<td>110.3</td>
<td>110.0</td>
<td>110.7</td>
<td>108.3</td>
</tr>
</tbody>
</table>

Source: Statistical Abstract of the United States

a. Draw a scatter plot of the poultry consumption and red meat consumption data. Is a linear regression appropriate?
b. Find the linear regression line that best fits the data. Use poultry consumption as your input variable.
c. What do the slope and y-intercept of your line mean in the context of the problem?
d. What is the estimate for red meat consumption if poultry consumption is 80 pounds per year?
e. Do you believe this is a correlation relationship or causation relationship? Explain.

20. Project: Medical Tests. The table below contains the fees charged by a Pacific Northwest medical laboratory to draw samples of a patient’s blood over the period from 2001 to 2009.

<table>
<thead>
<tr>
<th>Date</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
a. Draw a scatter plot of the data using year, $x$, as input and fee, $y$, as output.

b. Use a calculator to fit a linear equation to the data using linear regression.

c. Use the regression line to estimate the cost of a blood draw in 2010.

d. Predict the year in which the cost of a blood draw will reach $20.

e. Suggest in what year the blood draw would have been free (zero).

f. When data are generated by direct human action such as a management decision, as drawing fees are, they may contain hidden patterns. Find the change in fee, $\Delta y$, between each pair of years. Record it in a new row at the bottom of the table.

g. Using a calculator, fit a linear equation to the first five data points (years 1 to 5, when the change in fee per year was between 0.20 and 0.30). Explain why the equation is reasonable for those years.

h. Use the equation you found in part g to predict the drawing fee for 2006.

i. Fit a linear equation to the data points for 2005 to 2008, using the years 5 to 8 as the $x$-values and the fees as the $y$-values.

j. What is the meaning of the $y$-intercept in the equation you found in part i?

k. Use the equation in part i to predict the drawing fee for 2009.

l. Mentally determine the equation of the line passing through (8, 14.85) and (9, 14.85).

m. What economic factors might have influenced the lack of change from 2008 to 2009?

A sample should not be selected merely on the basis of convenience.

Research should be ethical as well as unbiased at every step, especially in taking the sample.

10.2 Descriptive Statistics: Averages

The mean, median, and mode are measures of central tendency. The mean is denoted by $\bar{x}$.

To find the mean of a set of numbers, add the numbers and divide by the number of numbers in the set.
To find the median, select the middle number when the numbers are arranged in numerical order (from smallest to largest or largest to smallest). If there is no single middle number, the median is the mean of the two middle numbers.

To find the mode, select the number or item appearing most frequently.

### 10.3 Variation and Summary

Measures of variation, or dispersion, include the range, five-number summary, quartiles, percentiles, and standard deviation. We illustrate statistics with box and whisker plots, bar graphs, and histograms.

The range is the difference between the largest and smallest numbers in a data set.

To draw a box and whisker plot, arrange the data in order from smallest to largest. Locate the median on a horizontal line with a scale appropriate to the data set. The quartiles are the midpoints of the first half and of the second half of the data. Locate the quartiles, $Q_1$ and $Q_3$, on the line. Draw a rectangle from $Q_1$ to $Q_3$. The whiskers extend out of the box to the minimum and maximum data points.

### Review Exercises

1. **Your neighbor complains that gasoline taxes are causing the prices at the pump to be high. Answering the questions below about the following sample data (from January of each year) on the components of gasoline prices will help you respond to your neighbor.**

   **What We Pay for in a Gallon of Regular Gasoline**

<table>
<thead>
<tr>
<th>Year (January)</th>
<th>Retail Price (dollars per gallon)</th>
<th>Refining (%)</th>
<th>Distribution &amp; Marketing (%)</th>
<th>Taxes (%)</th>
<th>Crude Oil (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>1.289</td>
<td>7.8</td>
<td>13.0</td>
<td>32.1</td>
<td>47.1</td>
</tr>
<tr>
<td>2001</td>
<td>1.447</td>
<td>17.8</td>
<td>10.4</td>
<td>29.2</td>
<td>42.7</td>
</tr>
<tr>
<td>2002</td>
<td>1.107</td>
<td>13.0</td>
<td>11.8</td>
<td>37.9</td>
<td>37.2</td>
</tr>
<tr>
<td>2003</td>
<td>1.458</td>
<td>11.5</td>
<td>10.3</td>
<td>28.8</td>
<td>49.4</td>
</tr>
<tr>
<td>2004</td>
<td>1.572</td>
<td>15.9</td>
<td>9.9</td>
<td>26.7</td>
<td>47.5</td>
</tr>
<tr>
<td>2005</td>
<td>1.831</td>
<td>17.7</td>
<td>7.3</td>
<td>24.0</td>
<td>50.9</td>
</tr>
<tr>
<td>2006</td>
<td>2.316</td>
<td>13.4</td>
<td>6.6</td>
<td>19.8</td>
<td>60.1</td>
</tr>
<tr>
<td>2007</td>
<td>2.240</td>
<td>10.6</td>
<td>15.2</td>
<td>20.3</td>
<td>53.9</td>
</tr>
<tr>
<td>2008</td>
<td>3.043</td>
<td>7.8</td>
<td>11.1</td>
<td>13.1</td>
<td>67.9</td>
</tr>
<tr>
<td>2009</td>
<td>1.788</td>
<td>13.4</td>
<td>10.7</td>
<td>22.3</td>
<td>53.6</td>
</tr>
<tr>
<td>2010</td>
<td>2.715</td>
<td>5.0</td>
<td>10.9</td>
<td>14.7</td>
<td>69.1</td>
</tr>
</tbody>
</table>

   - **a.** What kind of sample is shown in the table?
   - **b.** Which is the greatest factor in the price of gasoline: refining, distribution and marketing, taxes, or crude oil?
   - **c.** When were taxes the highest percentage of the cost of a gallon of gas? the lowest?
   - **d.** Calculate the tax per gallon for each year in dollars. When was it lowest? For example, for 2000, $0.321(1.289) = 0.414$.
   - **e.** Why was the tax percentage high in 2002? Why was it low in 2008?
   - **f.** Because of unrest in the Middle East, in 2011 gas prices soared to well over $3.50 per gallon. What happens to taxes as a percentage of the total cost of fuel in such a situation?
   - **g.** Is this an observational study or an experimental study?
   - **h.** Suppose you averaged a sample of gasoline prices from the table chosen by flipping a coin: heads, use the number; tails, don’t use the number. Would your sample be random?

For a set of $n$ numbers $x_1, x_2, x_3, \ldots, x_n$ with mean $\bar{x}$, drawn randomly from a population, the sample standard deviation is

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}}$$

The sample standard deviation, $s$, uses division by $n - 1$.

### 10.4 Linear Regression: Finding Equations from Data

If the plot of data appears to be linear, we may apply linear regression to fit a line to the data.

Statisticians use $y = ax + b$ instead of $y = mx + b$ to describe the regression line.

Warnings about linear regression:

- Linear regression may be used to estimate a data point within a set of data (interpolation), but caution should be exercised in using linear regression to estimate a point outside the set of data (extrapolation).
- The fact that linear regression yields a nice fit between sets of numbers does not imply that one set has any influence on the other: Correlation does not imply causation.
2. You are on the school board in a district that owns an old high school baseball stadium in bad repair. Three groups want to buy it: a historical preservation group with no money on hand but with neighborhood support, a YMCA with funds to tear the stadium down and build a variety of athletic and exercise facilities on the site, and a major supermarket chain. Describe how you would obtain community input on your decision. What population should you consider? What method might you use to obtain the input? What bias might you try to avoid?

3. Your mathematics teacher gives students the option of taking a test within the one-hour class time or taking it in a two-hour block in the Resource Center, but with a 5% deduction in grade. Students who have done the homework and attended class regularly can complete the test within an hour. Suppose a student wants to know whether it would be to his advantage to take the test within an hour. What kind of study would you undertake? What kind of sample would you use? What is the population? What bias might occur?

4. You work for a nonprofit agency and solicit donations over the telephone. You work after school on weeknights from 4 to 7 p.m. Your manager has decided to give prizes to those who collect the most money during the next three weeks. Is the contest fair? Why or why not?

5. The data below represent the number of chirps per minute by a sample of Rocky Mountain crickets.

<table>
<thead>
<tr>
<th>Chirps per Minute</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>50–54</td>
<td></td>
</tr>
<tr>
<td>55–59</td>
<td></td>
</tr>
<tr>
<td>60–64</td>
<td></td>
</tr>
<tr>
<td>65–69</td>
<td></td>
</tr>
<tr>
<td>70–74</td>
<td></td>
</tr>
<tr>
<td>75–79</td>
<td></td>
</tr>
</tbody>
</table>

a. Sort the data in ascending order.
b. Create a frequency chart for the data, using the following format.

c. Find the mean, median, and mode of the data.

6. The six geography majors graduating one year from a small college all enter the workforce. Their first-year salaries are listed below.

$34,000 $21,500 $19,750 $38,000 $22,000 $2,500,000

(The last salary belongs to a geography major who was also a basketball player and entered the NBA.)

a. Calculate the mean, median, and mode of the data.
b. The college uses the mean in its advertising to attract potential students. Do you think this is ethical? Why or why not? What measure of center would you use that might be more representative of the salary a graduate with a degree in geography might earn?

In Exercises 7 to 12, determine which measure of center would be most appropriate. Explain your answers.

7. Which ice cream is most popular among students at an ice cream shop.

8. What a typical home price is in a particular zip code. Most families in the zip code live below the poverty level, except for those in six homes in an exclusive gated community.

9. The amount each customer spent at a discount store on a given day

10. The most common male first name in the United States

11. The height of Army recruits in Minneapolis in 2010

12. The speed at which drivers drive on Interstate 40

Find the mean, median, and mode of each set of measurements in Exercises 13 to 16. Round to the nearest hundredth.

13. 4.2 grams, 4.3 grams, 4.3 grams

14. 31.7 cm, 31.8 cm, 31.6 cm, 31.8 cm

15. 6.9 miles, 6.9 miles, 6.8 miles, 7.0 miles, 6.8 miles, 6.8 miles

16. 45.0 mL, 44.0 mL, 45.5 mL, 45.0 mL, 44.5 mL

17. Staples sells large paper clips in groups of 5 boxes marked as containing 100 clips each. The actual count in 5 boxes is 96, 101, 98, 102, 102.

a. Find the mean.
b. Discuss the content from the point of view of the buyer.
c. Discuss the content from the point of view of Staples.
d. Would you accept the count as representative of all boxes?

18. Langche’s math instructor weights the term project and the final exam each as 2 and the two midterms each as 1. Langche gets 100 on the project, 85 on the final exam, and 95 and 75 on the midterms. What is his course average?
19. In traffic accident investigations, tire skid tests are used to find the coefficient of friction between tires and the road surface near an accident scene. An investigator measures the following tire skid marks, in feet, for a skid at 30 miles per hour:

Test 1: Left front, 50; right front, 49; left rear, 47; right rear, 48
Test 2: Left front, 47; right front, 50; left rear, 48; right rear, 51

a. Find the mean, range, and sample standard deviation \( s_x \) of the skid marks for each of the two tests. Round the standard deviation to two decimal places.
b. Find the coefficient of friction for each test with 
\[
f = \frac{S^2}{30D}
\]
where \( f \) is the coefficient of friction, \( S \) is the speed of the car making the tests, and \( D \) is the mean skid mark distance for the four tires.

20. The data below are for 6” Subway sandwiches.

<table>
<thead>
<tr>
<th>Sandwiches</th>
<th>Fat (g)</th>
<th>Cholesterol (mg)</th>
<th>Calories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subway club\textsuperscript{®}</td>
<td>5</td>
<td>26</td>
<td>312</td>
</tr>
<tr>
<td>Turkey breast &amp; ham</td>
<td>5</td>
<td>24</td>
<td>295</td>
</tr>
<tr>
<td>Veggie delite\textsuperscript{TM}</td>
<td>3</td>
<td>0</td>
<td>237</td>
</tr>
<tr>
<td>Turkey breast</td>
<td>4</td>
<td>19</td>
<td>289</td>
</tr>
<tr>
<td>Ham</td>
<td>5</td>
<td>28</td>
<td>302</td>
</tr>
<tr>
<td>Roast beef</td>
<td>5</td>
<td>20</td>
<td>303</td>
</tr>
<tr>
<td>Roasted chicken breast</td>
<td>6</td>
<td>48</td>
<td>348</td>
</tr>
</tbody>
</table>

**SUBWAY**\textsuperscript{®} regular 6” subs include bread, veggies, and meat. Addition of condiments or cheese alters nutrition content.

a. What are the mean, median, and mode for the grams of fat?
b. Draw a box and whisker plot for the milligrams of cholesterol.
c. What are the range, mean, and sample standard deviation for the calories?

21. The estimated numbers of U.S. children who were home schooled in the years from 1992 to 1997 are shown in the following table.

<table>
<thead>
<tr>
<th>Year (after 1992)</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>703,000</td>
</tr>
<tr>
<td>1</td>
<td>808,000</td>
</tr>
<tr>
<td>2</td>
<td>929,000</td>
</tr>
<tr>
<td>3</td>
<td>1,060,000</td>
</tr>
<tr>
<td>4</td>
<td>1,220,000</td>
</tr>
<tr>
<td>5</td>
<td>1,347,000</td>
</tr>
</tbody>
</table>

a. Write a function, \( N(t) \), that gives the best fit line to the number of children home schooled \( t \) years after 1992. Use linear regression to find this line.
b. According to the model you found in part a, how many children were home schooled in 1998?
c. According to the model you found in part a, when would 2,500,000 students be home schooled?

22. The following data set gives the per capita income in Washington state.

<table>
<thead>
<tr>
<th>Years after 2000</th>
<th>Personal Income ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>31,780</td>
</tr>
<tr>
<td>1</td>
<td>32,319</td>
</tr>
<tr>
<td>2</td>
<td>32,606</td>
</tr>
<tr>
<td>3</td>
<td>33,214</td>
</tr>
<tr>
<td>4</td>
<td>35,347</td>
</tr>
<tr>
<td>5</td>
<td>36,227</td>
</tr>
<tr>
<td>6</td>
<td>38,639</td>
</tr>
<tr>
<td>7</td>
<td>41,203</td>
</tr>
<tr>
<td>8</td>
<td>42,356</td>
</tr>
</tbody>
</table>

a. Draw a scatter plot of these data. Is a linear regression appropriate?
b. Find the linear regression line that best fits the data. Use years after 2000 as your input variable.
c. What do the slope and y-intercept of your line mean in the context of the problem?
d. What is the estimate for per capita personal income for Washington in 2012, according to your model?
e. When will per capita personal income exceed $50,000, according to your model?

23. **Project: Eyeglasses.** Those of us who wear eyeglasses have a source of data literally on our faces. Printed on glasses frames, generally between the lenses, is a pair of numbers separated by a square, such as 53\( \times \)17. These numbers, written in the table below with a comma, indicate the width of the lens and the width of the nose piece. On the temple pieces that extend over the ears is a three-digit number. This number indicates the length of the temple pieces in millimeters.

a. Find the mean of the first measurement in the first column for each data set (children, women, large men).
### Chapter 10 Review Exercises

#### 24. Project: Sparrow Migration

A classic study of sparrow migration was conducted by Margaret Morse Nice in Columbus, Ohio, between 1930 and 1935. Make observations about the data and support your statements with mathematics.

<table>
<thead>
<tr>
<th>Bird</th>
<th>1931</th>
<th>1932</th>
<th>Sex of Bird</th>
</tr>
</thead>
<tbody>
<tr>
<td>#2</td>
<td>Mar 20</td>
<td>Mar 1</td>
<td>male</td>
</tr>
<tr>
<td>#10</td>
<td>Apr 3</td>
<td>Mar 26</td>
<td>male</td>
</tr>
<tr>
<td>#23</td>
<td>Apr 3</td>
<td>Mar 21</td>
<td>male</td>
</tr>
<tr>
<td>#24</td>
<td>Mar 23</td>
<td>Feb 26</td>
<td>male</td>
</tr>
<tr>
<td>#47</td>
<td>Mar 10</td>
<td>Feb 26</td>
<td>male</td>
</tr>
<tr>
<td>#62</td>
<td>Mar 23</td>
<td>Mar 19</td>
<td>male</td>
</tr>
<tr>
<td>#64</td>
<td>Mar 30</td>
<td>Feb 26</td>
<td>male</td>
</tr>
<tr>
<td>#11</td>
<td>Apr 3</td>
<td>Mar 25</td>
<td>female</td>
</tr>
<tr>
<td>#14</td>
<td>Apr 3</td>
<td>Mar 29</td>
<td>female</td>
</tr>
<tr>
<td>#41</td>
<td>Mar 24</td>
<td>Mar 20</td>
<td>female</td>
</tr>
<tr>
<td>#46</td>
<td>Apr 3</td>
<td>Mar 25</td>
<td>female</td>
</tr>
<tr>
<td>#52</td>
<td>Apr 1</td>
<td>Mar 1</td>
<td>female</td>
</tr>
<tr>
<td>#58</td>
<td>Mar 24</td>
<td>Mar 3</td>
<td>female</td>
</tr>
<tr>
<td>#60</td>
<td>Apr 3</td>
<td>Mar 28</td>
<td>female</td>
</tr>
</tbody>
</table>

1. Give an example of a statistical study that uses observation.

2. Give an example of an experimental statistical study.

3. What bias must researchers avoid in mailing out surveys? How can they do this?

4. You are forming a large committee to interview the top three candidates for the next college president. You want the committee to be a sample representing all parts of campus: students, faculty, staff, and administration. Describe how you might assemble the committee.

What are the median and the mean of each set of data in Exercises 5 and 6?

5. Five ballpoint pens: $1.29, $1.29, $1.29, $1.29, $2.79

6. Five ballpoint pens: $1.69, $1.29, $1.19, $1.69, $2.09

7. Make some observations about the medians and means in Exercises 5 and 6 and what might have caused these results. Under what circumstances would each be a good description of the average cost of the pens?

8. The table below gives food energy and sodium content for a variety of dry cereals.

<table>
<thead>
<tr>
<th>Dry Cereal</th>
<th>Food Energy (calories)</th>
<th>Sodium Content (milligrams)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cap’n Crunch®</td>
<td>120</td>
<td>145</td>
</tr>
<tr>
<td>Froot Loops®</td>
<td>110</td>
<td>145</td>
</tr>
<tr>
<td>Super Golden Crisp®</td>
<td>105</td>
<td>25</td>
</tr>
<tr>
<td>Sugar Frosted Flakes®</td>
<td>110</td>
<td>230</td>
</tr>
<tr>
<td>Sugar Smacks®</td>
<td>105</td>
<td>75</td>
</tr>
<tr>
<td>Trix®</td>
<td>110</td>
<td>181</td>
</tr>
</tbody>
</table>

- a. Find the mean and sample standard deviation ($s_x$) for food energy.
- b. Find the median and make a box and whisker plot for sodium content.

9. Following are the winning times (in seconds) for the 500-meter event in speed skating at the past nine Olympic Games.

Men: 39.17, 38.03, 38.19, 36.45, 37.14, 36.33, 35.59, 34.42, 34.82

Women: 42.76, 41.78, 41.02, 39.10, 40.33, 39.25, 38.21, 37.30, 38.23

- a. Find the range of each set of data.
- b. Make a box and whisker plot for each set.
- c. Find the mean and sample standard deviation ($s_x$) for each set.
- d. What is the inter-quartile range (IQR) for each set of data? What does it mean in this problem setting?
- e. Draw a scatter plot for each set of data. Use 1 to 9 as inputs.
- f. Is either plot reasonably close to linear? Is a linear regression equation appropriate?
- g. Fit a linear equation to each plot. Why is the slope negative?

10. Suppose you were interested in buying a home for your family in a particular neighborhood in Providence, Rhode Island, near your new place of employment. Would you be most interested in knowing the mean, the median, or the mode of house prices? Explain your answer.

11. A poll was conducted by a news organization to predict the winner of a presidential campaign. Researchers called 100 college students in each state and asked who they were planning on voting for.

- a. What is the population?
- b. What is the population parameter?
- c. What is the sample?
- d. What is the sample statistic?
- e. What is the most remarkably bad type of bias in this survey?
Answers to Selected Odd-Numbered Exercises and Test

CHAPTER 10

Exercises 10.1

1. observational; Americans; 2010 randomly selected Americans; percent thinking we are less civil and percent thinking we are more likely to use vulgar language

3. experimental; population not clear; randomly selected individuals; percent doing task better in white room

5. experimental; students taking tests; students in this instructor’s class; average points

7. systematical sample; not valid: sample is likely to be biased by a periodic defect in process

9. grouped or stratified sample; if samples are proportional to population of students, then the sampling method is valid.

11. random sample (random numbers may not even give actual phone numbers); validity of method depends on nature of survey. If the survey is about phone service, this method is fine. If the survey is about the state of the economy, the method is not valid: many people no longer have listed phones or live in the local area.

13. a. all 3000 students; 25 self-selecting students

b. hours spent watching TV

c. average number of hours per day

d. convenience sample, not good

15. Bias may be economic; those eating lobster may be significantly wealthier.

17. Kiwi has a financial interest in the survey.

19. self-selecting bias (not to mention expensively duplicating work already done!)

21. Some data values are comparatively very large.

23. a. $723,000; $180,750; systematic

b. $928,000; $232,000; random

c. Answers vary.

d. 4.1%; answers vary.

Exercises 10.2

1. a. 85; 82.5; 80

b. 85; 82.5; 80

c. 52; 72.5; 0

d. 72.5; 85; 85

2. Students a and d are improving; b and e are not. Missing a quiz is not wise.

3. a. 477; 475; 475

b. 537; 525; 525 Mean apartment prices are lower in the smaller town by about $60.

5. Add the elements; divide by the number of elements.

7. No; for example, mean of 6 and 8 is 7 = mean of 5 and 9.

9. $A = h \cdot \frac{a + b}{2}$; area is height times average of lengths of parallel sides.

11. is not

13. Both describe the “middle” of something.

15. Find most-often-occurring number.

17. can increase mean; no effect on mode or median

19. when we want to reduce the effect of large or small data values

21. a. 18; 17; 17

b. It became more uniform across states, with fewer drivers age 15 or 16.

c. 1991: 17 and 17.03; 2002: 17 and 16.84; 2006: 17 and 17.01; 2010: 17 and 17.13

29. GRCC, 31; LCC, 29; BCC, 30

31. 0.85

33. 0.70

35. $=0.86$

37. not possible

Exercises 10.3

1. $26,000; $12,500; $92,000

3. $26,000; $27,500; $7500

5. $10,000; $13,000; data in thousands:

7. $23,750; $27,500; data in thousands:

9. $36,300

11. $3400

13. a. $7.5 + \frac{3}{4}$, or $7 \frac{3}{4}$ in.

b. $7.5 - \frac{3}{4}$, or $7 \frac{1}{4}$ in.

c. The $\frac{3}{4}$-in. variation is a range. No, the steps will not pass; the range from smallest to largest is $\frac{3}{4}$ in., which is greater than $\frac{3}{4}$ in.

15. a. 62.2; 13.6; 43, 50, 62.5, 75, 80

b. 52.3; 9.3; 41, 44, 51.5, 61, 66; cities have similar lows, but Atlanta has greater mean and high temperatures.
17. a. 77.6; 3.4; 73, 74.5, 77.5, 80.5, 82 b. 76.7; 5.9; 68, 71, 77.5, 82, 84; cities have nearly the same average temperature, but Miami has greater range and lower low temperatures.

19. a. 70.875; 74.5; 76; 16.44 b. 30

21. 6; 6.5; 7; 3, 4, 6.5, 8, 10; 2.02

23. for both data sets

25. The 07102 area has almost half its sales above $Q_3$ for the 07104 area. Looking at it another way, almost three quarters of the sales in area 07104 are below the median sales for 07102. The 07104 area had a narrower range of sales but wider IQR.

29. An outlier could stretch the whisker; a point in the IQR would have little or no effect.

Exercises 10.4

1. a. average slope = 0.6; (4, 4.99) gives $y = 0.6x + 2.59$

<table>
<thead>
<tr>
<th>$\Delta y$</th>
<th>$\Delta x$</th>
<th>Price, $y$ (dollars)</th>
<th>Volume, $x$ (cups)</th>
<th>Slope $\Delta y / \Delta x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.19</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.99</td>
<td>0.8</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>4</td>
<td>7.49</td>
<td>2.5</td>
<td>1</td>
<td>0.625</td>
</tr>
</tbody>
</table>

b. The equations describe lines that are almost parallel, with the 2010 prices shifted about $0.40 higher.

3. $y = 2.459x + 6.668; \$55.85; the product is now made in China.

5. $y = 0.0324x + 0.340; \$1.12; actual and predicted costs are similar.

7. a. no

b. yes; $y = 2.655x + 0.172$

c. area d. $\$0.38$ (the store price was $\$0.35$)
9. a. \( y = 3636x - 780 \)

\[ \text{Regular price (dollars)} \]

\[ \begin{array}{c|c|c|c|c|c|c}
\hline
\text{Weight (carat)} & 1/4 & 1/2 & 3/4 & 1 & 5/4 \\
\hline
\text{Regular price} & 2000 & 0 & 4000 & 0 & 2000 \\
\hline
\end{array} \]

b. \( y = 2545x - 547 \)

\[ \text{Price (dollars)} \]

\[ \begin{array}{c|c|c|c|c|c|c}
\hline
\text{Weight (carat)} & 1/4 & 1/2 & 3/4 & 1 & 5/4 \\
\hline
\text{Price} & 2000 & 0 & 4000 & 0 & 2000 \\
\hline
\end{array} \]

c. \( y = 0.7x - 1 \); the sale price is 70% of the regular price; the sale was a 30% discount.

11. a. \( y = 460.68x + 49,334.75 \)

b. Slope of 460.68 is the amount the population increases per year, y-intercept of 49,334.75 is what the model predicts for the population in 2000.

c. about 54,863 residents

d. in about 12.3 years, or sometime in 2012

13. a. yes  

b. \( y = 0.661x + 3.025 \)

c. Slope of 0.661 is the increase in lung cancer deaths per increase of 100 in per capita cigarette smoking rate. With no smoking we expect 3.025 deaths from lung cancer per 100,000 people.

d. 6.33  

e. 28.81

15. a. yes, as there is a general upward trend, but other models may be more appropriate

b. \( y = 496x + 2548.5 \)

c. Slope of 496 is the increase in public debt per year in billions of dollars; the y-intercept of 2548.5 is what the model predicts for 2000, in billions of dollars.

d. about $8500 billion  

e. in 2015

17. a. yes  

b. \( y = 9200x - 47,800 \)

c. $16,600. We hope you do not believe this estimate. Perhaps two of the friends were basketball players. It is coincidental that the data fit a line. A better sample might suggest a different result. This observational study provides no evidence of causation.

d. about 109 lb per year

e. This is a correlation relationship; it does not seem that poultry consumption directly causes a decline in red meat consumption. It may be that poultry stayed relatively cheap while red meat prices rose.
Review Exercises Chapter 10

1. a. systematic; data are recorded for January each year. b. crude oil c. 2002; 2008 d. $0.423, $0.420, $0.420, $0.439, $0.459, $0.455, $0.399, $0.399, $0.399; the tax per gallon was lower from 2008 to 2010. e. the price of gas was low; the price of gas was high f. the tax as a percentage drops g. observational h. yes

3. Possible answer: An experiment would be the preferable method. However, an observational study would be more practical. If the teacher allowed access to the data (with identifiers removed), you wouldn’t need to take a sample—you could use the entire set of data. If the teacher did not allow access, you might need to use a convenience sample, asking those you know who have previously taken the instructor’s course. The population might be all students who have ever taken the course with that particular instructor. Sources of bias: if you used a survey, students might not truthfully report their grades; weaker students might prefer to have a longer time to take a test, whereas stronger students might not want to take the deduction.

5. b. Frequencies are 3, 5, 8, 16, 7, 1 c. 64.625; 65.5; 65 7. mode 9. median 11. mean 13. 4.27; 4.3 15. 6.87; 6.85; 6.8 17. a. 99.8 b. and c. Answers vary. d. No; the range is 6 clips. You might expect more variance if you examined more boxes.

19. a. 48.5, 3, 1.29; 49, 4, 1.83 b. =0.62; =0.61 21. a. \(N(t) = 131.057t + 683.524\) b. 1,469,867 c. in 13 to 14 years from 1992, or 2005 or 2006

Chapter 10 Test

3. self-selection; they can do phone or e-mail follow-up to remind people to send in surveys or they can offer a reward for sending in the survey.

4. Use proportional sampling based on the relative proportions of groups on campus.

5. $1.29; $1.59 6. $1.69; $1.59 7. Answers vary.

8. a. 110; =5.5 b. 145

10. You probably would be interested in knowing the median house price. This average would not be affected by people buying very expensive homes, nor would it be affected by a small number of people buying fixer-uppers.

11. a. Americans who will vote in the upcoming election b. percent of voters who will vote for various candidates c. the 100 college students in the 50 states d. percent of those in the sample who will vote for various candidates e. selection bias; college students are not representative of general voters (and are not likely to vote).
CHAPTER 10

Activity 10.1

3. Answers will vary; here is one possibility. Select a random sample of size 6 on your calculator by entering randInt(1, 20, 6). Use the numbers of keys that correspond to the student numbers generated to calculate the sample statistic. Here are some possible results:

<table>
<thead>
<tr>
<th>Sample</th>
<th>Keys</th>
<th>Average Number of Keys in Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>{11, 13, 3, 10, 19, 7}</td>
<td>{10, 1, 3, 7, 5, 3}</td>
<td>4.83</td>
</tr>
<tr>
<td>{10, 13, 12, 19, 8, 15}</td>
<td>{7, 1, 3, 5, 6, 5}</td>
<td>4.5</td>
</tr>
<tr>
<td>{20, 12, 7, 6, 8, 2}</td>
<td>{3, 3, 3, 2, 6, 0}</td>
<td>2.83</td>
</tr>
<tr>
<td>{16, 9, 2, 13, 18, 1}</td>
<td>{3, 4, 0, 1, 8, 5}</td>
<td>3.5</td>
</tr>
<tr>
<td>{2, 7, 15, 14, 18, 4}</td>
<td>{0, 3, 5, 2, 8, 12}</td>
<td>5</td>
</tr>
</tbody>
</table>

4. and 5. The table below gives the answers for the samples in part 3.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Average Number of Keys in Sample</th>
<th>Percentage Error from the Actual Population Parameter (rounded to nearest percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>{11, 13, 3, 10, 19, 7}</td>
<td>4.83, larger</td>
<td>11%</td>
</tr>
<tr>
<td>{10, 13, 12, 19, 8, 15}</td>
<td>4.5, larger</td>
<td>3%</td>
</tr>
<tr>
<td>{20, 12, 7, 6, 8, 2}</td>
<td>2.83, smaller</td>
<td>35%</td>
</tr>
<tr>
<td>{16, 9, 2, 13, 18, 1}</td>
<td>3.5, smaller</td>
<td>20%</td>
</tr>
<tr>
<td>{2, 7, 15, 14, 18, 4}</td>
<td>5, larger</td>
<td>15%</td>
</tr>
</tbody>
</table>

Exercises 10.1

2. observational; first-year college students; 235,812 first-year students; percent with A average
4. observational; sparrows in all count areas; sparrows in polluted count areas; how many more
6. random sample; if samples are proportional to population of students, then the sampling method is valid.
8. convenience sample; will get opinions but may not represent all citizens of your state

10. convenience sample; not valid unless people who frequent the playground area (such as parents of young children) are your target market
12. a. all students; 100 dorm students
   b. exercise time from surveys
   c. average exercise time
   d. Unless all students at this college live in the dorms, parts of the student population are ignored.

27. a. The population is the set of airline travelers. The population parameter is their frustration level. You will need to take an unbiased (fair) sample of travelers and do a survey (perhaps a survey by groups in randomly selected planes after travelers board the planes, as experienced by one author). The results from this survey are the raw data. You then might calculate some type of statistic (an average or proportion) that mathematically illustrates travelers’ frustration level. From this you could make an inference about the population parameter (percentage who are frustrated, for example) and draw a conclusion.
   b. Answers vary.
   c. Before and after making changes to security processes, do one or more of the following: Count the number of negative interactions between TSA employees and passengers. Count the number of complaints filed with the airport. Time the wait time for randomly selected passengers.

Activity 10.2

2. The sorted data are shown below.

<table>
<thead>
<tr>
<th>Min</th>
<th>54</th>
<th>54</th>
<th>60</th>
<th>60</th>
<th>60</th>
<th>60</th>
<th>60</th>
<th>60</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>66</td>
<td>66</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>72</td>
<td>72</td>
<td>78</td>
<td>84</td>
<td>84</td>
<td>84</td>
<td>90</td>
<td>96</td>
<td>156</td>
</tr>
</tbody>
</table>

Activity 10.3

1. 42, 43, 46, 46, 47, 47, 48, 49, 49, 50, 51, 51, 51, 51, 52, 52, 54, 54, 54, 54, 54, 55, 55, 55, 56, 56, 57, 57, 57, 57, 58, 60, 61, 61, 62, 64, 64, 65, 68, 69
4. 51 54.5 57.5

Min 42 Med 69 Max

12/9/11 9:20 AM
**Exercises 10.3**

6. Data in thousands:

![Q1 and Q3 values]

8. Data in thousands:

![Q1, Med, Q3 values]

22. a. 

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>70–79</td>
<td>2</td>
</tr>
<tr>
<td>80–89</td>
<td>2</td>
</tr>
<tr>
<td>90–99</td>
<td>3</td>
</tr>
<tr>
<td>100–109</td>
<td>5</td>
</tr>
<tr>
<td>110–119</td>
<td>2</td>
</tr>
<tr>
<td>120–129</td>
<td>2</td>
</tr>
<tr>
<td>130–139</td>
<td>2</td>
</tr>
<tr>
<td>140–149</td>
<td>2</td>
</tr>
</tbody>
</table>

b. The composition of the 1-cent piece changed in 1982. The pre-1982 coins are 95% copper and 5% zinc. The post-1982 coins are 97.6% zinc and 2.4% copper. The post-1982 coins are lighter because zinc is less dense than copper. For 1981 and before, \( \bar{x} = 3.110, s = 0.0623, n = 17 \); for 1982 and after, \( \bar{x} = 2.521, s = 0.04738, n = 10 \)

35. b.

![Box plot example]

**Activity 10.4**

1. 

2. 

<table>
<thead>
<tr>
<th>( \Delta x )</th>
<th>Volume, ( x ) (quarts)</th>
<th>Price, ( y ) (dollars)</th>
<th>( \Delta y )</th>
<th>Slope ( \Delta y/\Delta x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>10</td>
<td>2</td>
<td>0.67</td>
</tr>
</tbody>
</table>

4. Table contains (0, 5), (17, 10.99), (20, 11.89). \( y = 0.347x + 5.011; \$13.34; \) the product is now made elsewhere, or the economy or competition kept the price down.

6. Linear equation doesn’t fit very well, but if we fit the data to the line \( y = 0.032x + 1.023 \), we get \$1.80 for 2010, which matches well. Perhaps vegetable prices vary with supply of product. All are still grown in the United States.
Chapter 10 Review Exercises

2. A systematic sample could be chosen from a population that lives within 3 miles of the stadium, and a survey could be given to those selected. Avoid wording bias. Avoid bias that will arise if only those closest to the stadium respond.

4. Possible answer: not fair. You may work fewer hours than other employees. Daytime employees might not get a chance to talk with breadwinners. You would be likely to upset people by calling during dinner time.
22. a.

The median width of the nose piece is 17 mm for women and 16.5 mm for children. Half the women's frames in this sample are larger than 17 mm; half the children's are smaller than 16.5 mm. For women, \( m = Q_3 = 17 \) mm; right edge of box is 17 for both. The men's \( Q_1 \) is 18 and indicates that three fourths of the data are greater than or equal to 18, the maximum for both children's and women's nose pieces.

d. The median lens width for unisex glasses is the same as \( Q_3 \) for women's glasses, but their maximum lens width equals the minimum lens width for large men's glasses. A box plot of nose distances for unisex glasses is nearly identical to those for children and women! In fact, the median nose distance for unisex glasses is 1 less than the median for children's. For the unisex temples, \( Q_1 = m = Q_3 = 135 \), so unisex temples are longer on average than women's temples, which had \( \text{Min} = Q_1 = m = 130 \). The unisex glasses might work for men wanting smaller glasses than the large men's and with narrow nose distances (on average).

e. The nose width doesn’t change as much as the length of the temple.

23. b. The median width of the nose piece is 17 mm for women and 16.5 mm for children. Half the women's frames in this sample are larger than 17 mm; half the children's are smaller than 16.5 mm. For women, \( m = Q_3 = 17 \) mm; right edge of box is 17 for both. The men's \( Q_1 \) is 18 and indicates that three fourths of the data are greater than or equal to 18, the maximum for both children's and women's nose pieces.

c.

24. Male sparrows arrive in Columbus, Ohio, earlier in the year than females. Change the dates to days from Jan. 1 and average the dates, remembering that 1932 was a leap year. On average, male birds arrived before females (1931, 83.9 days into year compared to 89.9 days; 1932, 68.3 days into year compared to 78.7 days), and both sexes arrived earlier in 1932 than in 1931.