1

CHAPTER

INTRODUCTION TO STATISTICS

In 2008, the population of New Orleans, Louisiana grew faster than any other large city in the United States. Despite the increase, the population of 311,853 was still well below the pre-Hurricane Katrina population of 484,674.

WHERE YOU'VE BEEN

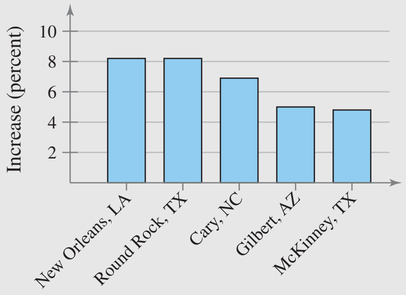
You are already familiar with many of the practices of statistics, such as taking surveys, collecting data, and describing populations. What you may not know is that collecting accurate statistical data is often difficult and costly. Consider, for instance, the monumental task of counting and describing the entire population of the United States. If you were in charge of such a census, how would you do it? How would you ensure that your results are accurate? These and many more concerns are the responsibility of the United States Census Bureau, which conducts the census every decade.

WHERE YOU'RE GOING

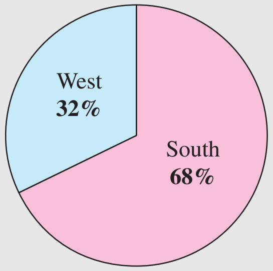
In Chapter 1, you will be introduced to the basic concepts and goals of statistics. For instance, statistics were used to construct the following graphs, which show the fastest growing U.S. cities (population over 100,000) in 2008 by percent increase in population, U.S. cities with the largest numerical increases in population, and the regions where the cities are located.

For the 2010 Census, the Census Bureau sent short forms to every household. Short forms ask all members of every household such things as their gender, age, race, and ethnicity. Previously, a long form, which covered additional topics, was sent to about 17% of the population. But for the first time since 1940, the long form is being replaced by the American Community Survey, which will survey about 3 million households a year throughout the decade. These 3 million households will form a sample. In this course, you will learn how the data collected from a sample are used to infer characteristics about the entire population.

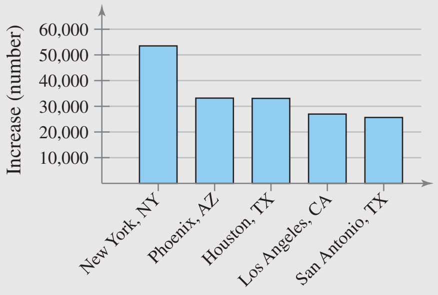
Fastest Growing U.S. Cities (Population over 100,000)



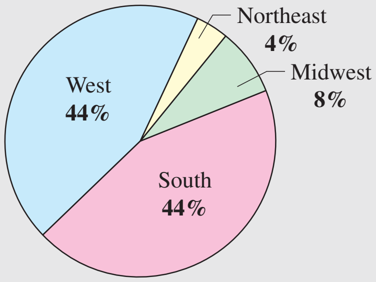
Location of the 25 Fastest Growing U.S. Cities



U.S. Cities with Largest Numerical Increases



Location of the 25 U.S. Cities with Largest Numerical Increases



1.1 An Overview of Statistics

WHAT YOU SHOULD LEARN

- The definition of statistics
- How to distinguish between a population and a sample and between a parameter and a statistic
- How to distinguish between descriptive statistics and inferential statistics

A Definition of Statistics ▸ Data Sets ▸ Branches of Statistics

A DEFINITION OF STATISTICS

As you begin this course, you may wonder: *What is statistics? Why should I study statistics? How can studying statistics help me in my profession?* Almost every day you are exposed to statistics. For instance, consider the following.

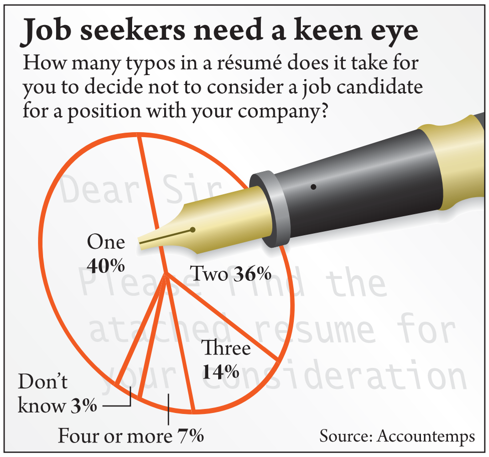
- "The number of Americans with diabetes will nearly double in the next 25 years." *(Source: Diabetes Care)*
- "The NRF expects holiday sales to decline 1% versus a 3.4% drop in holiday sales the previous year." *(Source: National Retail Federation)*
- "EIA projects total U.S. natural gas consumption will decline by 2.6 percent in 2009 and increase by 0.5 percent in 2010." *(Source: Energy Information Administration)*

The three statements you just read are based on the collection of *data*.

DEFINITION

Data consist of information coming from observations, counts, measurements, or responses.

Sometimes data are presented graphically. If you have ever read *USA TODAY*, you have certainly seen one of that newspaper's most popular features, *USA TODAY Snapshots*. Graphics such as this present information in a way that is easy to understand.



Source: Accountemps

The use of statistics dates back to census taking in ancient Babylonia, Egypt, and later in the Roman Empire, when data were collected about matters concerning the state, such as births and deaths. In fact, the word *statistics* is derived from the Latin word *status*, meaning "state." So, what is statistics?

DEFINITION

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

DATA SETS

There are two types of data sets you will use when studying statistics. These data sets are called *populations* and *samples*.

INSIGHT

A census consists of data from an entire population. But, unless a population is small, it is usually impractical to obtain all the population data. In most studies, information must be obtained from a sample.

DEFINITION

A **population** is the collection of *all* outcomes, responses, measurements, or counts that are of interest.

A **sample** is a subset, or part, of a population.

A sample should be representative of a population so that sample data can be used to form conclusions about that population. Sample data must be collected using an appropriate method, such as *random sampling*. (You will learn more about random sampling in Section 1.3.) If they are not collected using an appropriate method, the data are of no value.

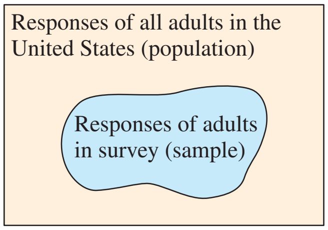
EXAMPLE 1

Identifying Data Sets

In a recent survey, 1500 adults in the United States were asked if they thought there was solid evidence of global warming. Eight hundred fifty-five of the adults said yes. Identify the population and the sample. Describe the sample data set. *(Adapted from Pew Research Center)*

Solution

The population consists of the responses of all adults in the United States, and the sample consists of the responses of the 1500 adults in the United States in the survey. The sample is a subset of the responses of all adults in the United States. The sample data set consists of 855 yes's and 645 no's.


Responses of all adults in the United States (population)
Responses of adults in survey (sample)

Try It Yourself 1

The U.S. Department of Energy conducts weekly surveys of approximately 900 gasoline stations to determine the average price per gallon of regular gasoline. On January 11, 2010, the average price was $2.75 per gallon. Identify the population and the sample. Describe the sample data set. *(Source: Energy Information Administration)*

a. Identify the *population* and the *sample*.
b. What does the sample data set consist of? *Answer: Page A30*

Whether a data set is a population or a sample usually depends on the context of the real-life situation. For instance, in Example 1, the population was the set of responses of all adults in the United States. Depending on the purpose of the survey, the population could have been the set of responses of all adults who live in California or who have cellular phones or who read a particular magazine.

Two important terms that are used throughout this course are *parameter* and *statistic*.

DEFINITION

A **parameter** is a numerical description of a *population* characteristic.

A **statistic** is a numerical description of a *sample* characteristic.

STUDY TIP

The terms *parameter* and *statistic* are easy to remember if you use the mnemonic device of matching the first letters in ***p**opulation **p**arameter* and the first letters in ***s**ample **s**tatistic*.

It is important to note that a sample statistic can differ from sample to sample whereas a population parameter is constant for a population.

EXAMPLE 2

Distinguishing Between a Parameter and a Statistic

Decide whether the numerical value describes a population parameter or a sample statistic. Explain your reasoning.

1. A recent survey of 200 college career centers reported that the average starting salary for petroleum engineering majors is $83,121. *(Source: National Association of Colleges and Employers)*
2. The 2182 students who accepted admission offers to Northwestern University in 2009 have an average SAT score of 1442. *(Source: Northwestern University)*
3. In a random check of a sample of retail stores, the Food and Drug Administration found that 34% of the stores were not storing fish at the proper temperature.

Solution

1. Because the average of $83,121 is based on a subset of the population, it is a sample statistic.
2. Because the SAT score of 1442 is based on all the students who accepted admission offers in 2009, it is a population parameter.
3. Because the percent of 34% is based on a subset of the population, it is a sample statistic.

Try It Yourself 2

In 2009, Major League Baseball teams spent a total of $2,655,395,194 on players' salaries. Does this numerical value describe a population parameter or a sample statistic? *(Source: USA Today)*

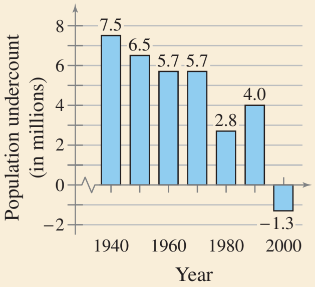
a. Decide whether the numerical value is from a *population* or a *sample*.
b. Specify whether the numerical value is a *parameter* or a *statistic*.

Answer: Page A30

PICTURING THE WORLD

How accurate is the U.S. census? According to a post-census evaluation conducted by the Census Bureau, the 1990 census undercounted the U.S. population by an estimated 4.0 million people. The 1990 census was the first census since at least 1940 to be less accurate than its predecessor. Notice that the undercount for the 2000 census was −1.3 million people. This means that the 2000 census *overcounted* the U.S. population by 1.3 million people.

U.S. Census Undercount



What are some difficulties in collecting population data?

In this course, you will see how the use of statistics can help you make informed decisions that affect your life. Consider the census that the U.S. government takes every decade. When taking the census, the Census Bureau attempts to contact everyone living in the United States. Although it is impossible to count everyone, it is important that the census be as accurate as it can be, because public officials make many decisions based on the census information. Data collected in the 2010 census will determine how to assign congressional seats and how to distribute public funds.

BRANCHES OF STATISTICS

The study of statistics has two major branches: *descriptive statistics* and *inferential statistics*.

DEFINITION

Descriptive statistics is the branch of statistics that involves the organization, summarization, and display of data.

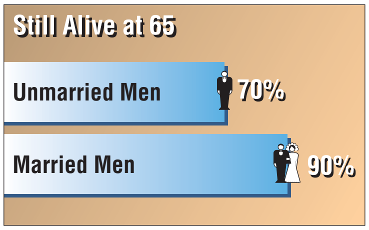
Inferential statistics is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

EXAMPLE 3

Descriptive and Inferential Statistics

Decide which part of the study represents the descriptive branch of statistics. What conclusions might be drawn from the study using inferential statistics?

1. A large sample of men, aged 48, was studied for 18 years. For unmarried men, approximately 70% were alive at age 65. For married men, 90% were alive at age 65. *(Source: The Journal of Family Issues)*



2. In a sample of Wall Street analysts, the percentage who incorrectly forecasted high-tech earnings in a recent year was 44%. *(Source: Bloomberg News)*

Solution

1. Descriptive statistics involves statements such as “For unmarried men, approximately 70% were alive at age 65” and “For married men, 90% were alive at 65.” A possible inference drawn from the study is that being married is associated with a longer life for men.

2. The part of this study that represents the descriptive branch of statistics involves the statement “the percentage [of Wall Street analysts] who incorrectly forecasted high-tech earnings in a recent year was 44%.” A possible inference drawn from the study is that the stock market is difficult to forecast, even for professionals.

Try It Yourself 3

A survey conducted among 1017 men and women by Opinion Research Corporation International found that 76% of women and 60% of men had a physical examination within the previous year. *(Source: Men’s Health)*

a. Identify the *descriptive* aspect of the survey.
b. What *inferences* could be drawn from this survey? *Answer: Page A30*

Throughout this course you will see applications of both branches. A major theme in this course will be how to use sample statistics to make inferences about unknown population parameters.

1.1 EXERCISES

BUILDING BASIC SKILLS AND VOCABULARY

1. How is a sample related to a population?

2. Why is a sample used more often than a population?

3. What is the difference between a parameter and a statistic?

4. What are the two main branches of statistics?

True or False? *In Exercises 5–10, determine whether the statement is true or false. If it is false, rewrite it as a true statement.*

5. A statistic is a measure that describes a population characteristic.

6. A sample is a subset of a population.

7. It is impossible for the Census Bureau to obtain all the census data about the population of the United States.

8. Inferential statistics involves using a population to draw a conclusion about a corresponding sample.

9. A population is the collection of some outcomes, responses, measurements, or counts that are of interest.

10. A sample statistic will not change from sample to sample.

Classifying a Data Set *In Exercises 11–20, determine whether the data set is a population or a sample. Explain your reasoning.*

11. The height of each player on a school's basketball team

12. The amount of energy collected from every wind turbine on a wind farm

13. A survey of 500 spectators from a stadium with 42,000 spectators

14. The annual salary of each pharmacist at a pharmacy

15. The cholesterol levels of 20 patients in a hospital with 100 patients

16. The number of televisions in each U.S. household

17. The final score of each golfer in a tournament

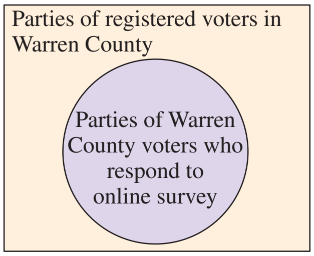
18. The age of every third person entering a clothing store

19. The political party of every U.S. president

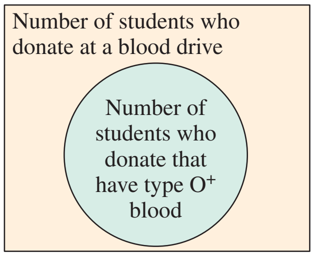
20. The soil contamination levels at 10 locations near a landfill

Graphical Analysis *In Exercises 21–24, use the Venn diagram to identify the population and the sample.*

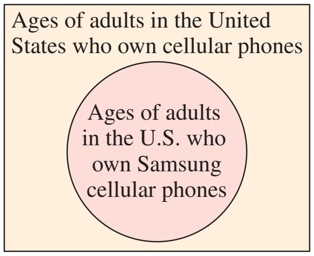
21.


Parties of registered voters in Warren County
Parties of Warren County voters who respond to online survey

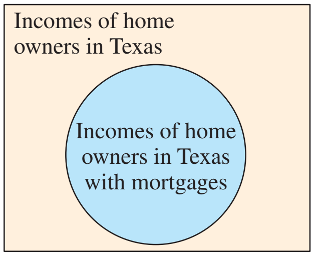
22.


Number of students who donate at a blood drive
Number of students who donate that have type O⁺ blood

23.


Ages of adults in the United States who own cellular phones
Ages of adults in the U.S. who own Samsung cellular phones

24.


Incomes of home owners in Texas
Incomes of home owners in Texas with mortgages

USING AND INTERPRETING CONCEPTS

Identifying Populations and Samples *In Exercises 25–34, identify the population and the sample.*

25. A survey of 1000 U.S. adults found that 59% think buying a home is the best investment a family can make. *(Source: Rasmussen Reports)*

26. A study of 33,043 infants in Italy was conducted to find a link between a heart rhythm abnormality and sudden infant death syndrome. *(Source: New England Journal of Medicine)*

27. A survey of 1442 U.S. adults found that 36% received an influenza vaccine for the current flu season. *(Source: Zogby International)*

28. A survey of 1600 people found that 76% plan on using the Microsoft Windows 7™ operating system at their businesses. *(Source: Information Technology Intelligence Corporation and Sunbelt Software)*

29. A survey of 800 registered voters found that 50% think economic stimulus is the most important issue to consider when voting for Congress. *(Source: Diageo/Hotline Poll)*

30. A survey of 496 students at a college found that 10% planned on traveling out of the country during spring break.

31. A survey of 546 U.S. women found that more than 56% are the primary investors in their households. *(Adapted from Roper Starch Worldwide for Intuit)*

32. A survey of 791 vacationers from the United States found that they planned on spending at least $2000 for their next vacation.

33. A magazine mails questionnaires to each company in Fortune magazine's top 100 best companies to work for and receives responses from 85 of them.

34. At the end of the day, a quality control inspector selects 20 light bulbs from the day's production and tests them.

Distinguishing Between a Parameter and a Statistic *In Exercises 35–42, determine whether the numerical value is a parameter or a statistic. Explain your reasoning.*

35. The average annual salary for 35 of a company's 1200 accountants is $68,000.

36. In a survey of a sample of high school students, 43% said that their mothers had taught them the most about managing money. *(Source: Harris Poll for Girls Incorporated)*

37. Sixty-two of the 97 passengers aboard the Hindenburg airship survived its explosion.

38. In January 2010, 52% of the governors of the 50 states in the United States were Democrats.

39. In a survey of 300 computer users, 8% said their computers had malfunctions that needed to be repaired by service technicians.

40. In a recent year, the interest category for 12% of all new magazines was sports. *(Source: Oxbridge Communications)*

41. In a recent survey of 2000 people, 44% said China is the world's leading economic power. *(Source: Pew Research Center)*

42. In a recent year, the average math scores for all graduates on the ACT was 21.0. *(Source: ACT, Inc.)*

43. Which part of the survey described in Exercise 31 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

44. Which part of the survey described in Exercise 32 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

EXTENDING CONCEPTS

45. Identifying Data Sets in Articles Find a newspaper or magazine article that describes a survey.

(a) Identify the sample used in the survey.

(b) What is the sample's population?

(c) Make an inference based on the results of the survey.

46. Sleep Deprivation In a recent study, volunteers who had 8 hours of sleep were three times more likely to answer questions correctly on a math test than were sleep-deprived participants. *(Source: CBS News)*

(a) Identify the sample used in the study.

(b) What is the sample's population?

(c) Which part of the study represents the descriptive branch of statistics?

(d) Make an inference based on the results of the study.

47. Living in Florida A study shows that senior citizens who live in Florida have better memories than senior citizens who do not live in Florida.

(a) Make an inference based on the results of this study.

(b) What is wrong with this type of reasoning?

48. Increase in Obesity Rates A study shows that the obesity rate among boys ages 2 to 19 has increased over the past several years. *(Source: Washington Post)*

(a) Make an inference based on the results of this study.

(b) What is wrong with this type of reasoning?

49. Writing Write an essay about the importance of statistics for one of the following.

- A study on the effectiveness of a new drug
- An analysis of a manufacturing process
- Making conclusions about voter opinions using surveys

1.2 Data Classification

WHAT YOU SHOULD LEARN

- How to distinguish between qualitative data and quantitative data
- How to classify data with respect to the four levels of measurement: nominal, ordinal, interval, and ratio

Types of Data ▸ Levels of Measurement

▸ TYPES OF DATA

When doing a study, it is important to know the kind of data involved. The nature of the data you are working with will determine which statistical procedures can be used. In this section, you will learn how to classify data by type and by level of measurement. Data sets can consist of two types of data: *qualitative data* and *quantitative data*.

DEFINITION

Qualitative data consist of attributes, labels, or nonnumerical entries.

Quantitative data consist of numerical measurements or counts.

EXAMPLE 1

▸ Classifying Data by Type

The suggested retail prices of several Ford vehicles are shown in the table. Which data are qualitative data and which are quantitative data? Explain your reasoning. *(Source: Ford Motor Company)*

Model	Suggested retail price
Focus Sedan	$15,995
Fusion	$19,270
Mustang	$20,995
Edge	$26,920
Flex	$28,495
Escape Hybrid	$32,260
Expedition	$35,085
F-450	$44,145

▸ Solution

The information shown in the table can be separated into two data sets. One data set contains the names of vehicle models, and the other contains the suggested retail prices of vehicle models. The names are nonnumerical entries, so these are qualitative data. The suggested retail prices are numerical entries, so these are quantitative data.

▸ Try It Yourself 1

The populations of several U.S. cities are shown in the table. Which data are qualitative data and which are quantitative data? *(Source: U.S. Census Bureau)*

City	Population
Baltimore, MD	636,919
Jacksonville, FL	807,815
Memphis, TN	669,651
Pasadena, CA	143,080
San Antonio, TX	1,351,305
Seattle, WA	598,541

a. *Identify* the two data sets.
b. Decide whether each data set consists of *numerical* or *nonnumerical* entries.
c. Specify the *qualitative* data and the *quantitative* data. *Answer: Page A30*

LEVELS OF MEASUREMENT

Another characteristic of data is its level of measurement. The level of measurement determines which statistical calculations are meaningful. The four levels of measurement, in order from lowest to highest, are *nominal*, *ordinal*, *interval*, and *ratio*.

DEFINITION

Data at the **nominal level of measurement** are qualitative only. Data at this level are categorized using names, labels, or qualities. No mathematical computations can be made at this level.

Data at the **ordinal level of measurement** are qualitative or quantitative. Data at this level can be arranged in order, or ranked, but differences between data entries are not meaningful.

When numbers are at the nominal level of measurement, they simply represent a label. Examples of numbers used as labels include Social Security numbers and numbers on sports jerseys. For instance, it would not make sense to add the numbers on the players' jerseys for the Chicago Bears.

EXAMPLE 2

Classifying Data by Level

Two data sets are shown. Which data set consists of data at the nominal level? Which data set consists of data at the ordinal level? Explain your reasoning. *(Source: The Nielsen Company)*

Top Five TV Programs (from 5/4/09 to 5/10/09)
1. American Idol–Wednesday
2. American Idol–Tuesday
3. Dancing with the Stars
4. NCIS
5. The Mentalist

Network Affiliates in Pittsburgh, PA	
WTAE	(ABC)
WPXI	(NBC)
KDKA	(CBS)
WPGH	(FOX)

Solution

The first data set lists the ranks of five TV programs. The data set consists of the ranks 1, 2, 3, 4, and 5. Because the ranks can be listed in order, these data are at the ordinal level. Note that the difference between a rank of 1 and 5 has no mathematical meaning. The second data set consists of the call letters of each network affiliate in Pittsburgh. The call letters are simply the names of network affiliates, so these data are at the nominal level.

Try It Yourself 2

Consider the following data sets. For each data set, decide whether the data are at the nominal level or at the ordinal level.

1. The final standings for the Pacific Division of the National Basketball Association

2. A collection of phone numbers

a. *Identify* what each data set represents.
b. Specify the *level of measurement* and justify your answer.

Answer: Page A30

PICTURING THE WORLD

In 2009, Forbes Magazine chose the 75 best business schools in the United States. Forbes based their rankings on the return on investment achieved by the graduates from the class of 2004. Graduates of the top five M.B.A. programs typically earn more than $200,000 within five years. (Source: Forbes)

Forbes Top Five U.S. Business Schools
1. Stanford
2. Dartmouth
3. Harvard
4. Chicago
5. Pennsylvania

In this list, what is the level of measurement?

The two highest levels of measurement consist of quantitative data only.

DEFINITION

Data at the **interval level of measurement** can be ordered, and meaningful differences between data entries can be calculated. At the interval level, a zero entry simply represents a position on a scale; the entry is not an inherent zero.

Data at the **ratio level of measurement** are similar to data at the interval level, with the added property that a zero entry is an inherent zero. A ratio of two data values can be formed so that one data value can be meaningfully expressed as a multiple of another.

An *inherent zero* is a zero that implies "none." For instance, the amount of money you have in a savings account could be zero dollars. In this case, the zero represents no money; it is an inherent zero. On the other hand, a temperature of 0°C does not represent a condition in which no heat is present. The 0°C temperature is simply a position on the Celsius scale; it is not an inherent zero.

To distinguish between data at the interval level and at the ratio level, determine whether the expression "twice as much" has any meaning in the context of the data. For instance, \$2 is twice as much as \$1, so these data are at the ratio level. On the other hand, 2°C is not twice as warm as 1°C, so these data are at the interval level.

New York Yankees' World Series Victories (Years)
1923, 1927, 1928, 1932, 1936, 1937, 1938, 1939, 1941, 1943, 1947, 1949, 1950, 1951, 1952, 1953, 1956, 1958, 1961, 1962, 1977, 1978, 1996, 1998, 1999, 2000, 2009

2009 American League Home Run Totals (by Team)	
Baltimore	160
Boston	212
Chicago	184
Cleveland	161
Detroit	183
Kansas City	144
Los Angeles	173
Minnesota	172
New York	244
Oakland	135
Seattle	160
Tampa Bay	199
Texas	224
Toronto	209

EXAMPLE 3

▸ Classifying Data by Level

Two data sets are shown at the left. Which data set consists of data at the interval level? Which data set consists of data at the ratio level? Explain your reasoning. *(Source: Major League Baseball)*

▸ Solution

Both of these data sets contain quantitative data. Consider the dates of the Yankees' World Series victories. It makes sense to find differences between specific dates. For instance, the time between the Yankees' first and last World Series victories is

$$2009 - 1923 = 86 \text{ years.}$$

But it does not make sense to say that one year is a multiple of another. So, these data are at the interval level. However, using the home run totals, you can find differences *and* write ratios. From the data, you can see that Texas hit 63 more home runs than Cleveland hit and that New York hit about 1.5 times as many home runs as Seattle hit. So, these data are at the ratio level.

▸ Try It Yourself 3

Decide whether the data are at the interval level or at the ratio level.

1. The body temperatures (in degrees Fahrenheit) of an athlete during an exercise session

2. The heart rates (in beats per minute) of an athlete during an exercise session

a. *Identify* what each data set represents.
b. Specify the *level of measurement* and justify your answer.

Answer: Page A30

The following tables summarize which operations are meaningful at each of the four levels of measurement. When identifying a data set's level of measurement, use the highest level that applies.

Level of measurement	**Put data in categories**	**Arrange data in order**	**Subtract data values**	**Determine if one data value is a multiple of another**
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Summary of Four Levels of Measurement

	Example of a Data Set	**Meaningful Calculations**
Nominal Level (Qualitative data)	*Types of Shows Televised by a Network* Comedy Drama Reality Shows Sports Documentaries Cooking Soap Operas Talk Shows	*Put in a category.* For instance, a show televised by the network could be put into one of the eight categories shown.
Ordinal Level (Qualitative or quantitative data)	*Motion Picture Association of America Ratings Description* G General Audiences PG Parental Guidance Suggested PG-13 Parents Strongly Cautioned R Restricted NC-17 No One Under 17 Admitted	Put in a category and *put in order*. For instance, a PG rating has a stronger restriction than a G rating.
Interval Level (Quantitative data)	*Average Monthly Temperatures (in degrees Fahrenheit) for Denver, CO* Jan 29.2 Feb 33.2 Mar 39.6 Apr 47.6 May 57.2 Jun 67.6 Jul 73.4 Aug 71.7 Sep 62.4 Oct 51.0 Nov 37.5 Dec 30.3 *(Source: National Climatic Data Center)*	Put in a category, put in order, and *find differences between values.* For instance, $57.2 - 47.6 = 9.6$°F. So, May is 9.6° warmer than April.
Ratio Level (Quantitative data)	*Average Monthly Precipitation (in inches) for Orlando, FL* Jan 2.4 Feb 2.4 Mar 3.5 Apr 2.4 May 3.7 Jun 7.4 Jul 7.2 Aug 6.3 Sep 5.8 Oct 2.7 Nov 2.3 Dec 2.3 *(Source: National Climatic Data Center)*	Put in a category, put in order, find differences between values, and *find ratios of values.* For instance, $\frac{7.4}{3.7} = 2$. So, there is twice as much rain in June as in May.

1.2 EXERCISES

BUILDING BASIC SKILLS AND VOCABULARY

1. Name each level of measurement for which data can be qualitative.

2. Name each level of measurement for which data can be quantitative.

True or False? *In Exercises 3–6, determine whether the statement is true or false. If it is false, rewrite it as a true statement.*

3. Data at the ordinal level are quantitative only.

4. For data at the interval level, you cannot calculate meaningful differences between data entries.

5. More types of calculations can be performed with data at the nominal level than with data at the interval level.

6. Data at the ratio level cannot be put in order.

USING AND INTERPRETING CONCEPTS

Classifying Data by Type *In Exercises 7–18, determine whether the data are qualitative or quantitative. Explain your reasoning.*

7. telephone numbers in a directory

8. heights of hot air balloons

9. body temperatures of patients

10. eye colors of models

11. lengths of songs on MP3 player

12. carrying capacities of pickups

13. player numbers for a soccer team

14. student ID numbers

15. weights of infants at a hospital

16. species of trees in a forest

17. responses on an opinion poll

18. wait times at a grocery store

Classifying Data by Level *In Exercises 19–24, determine whether the data are qualitative or quantitative, and identify the data set's level of measurement. Explain your reasoning.*

19. Football The top five teams in the final college football poll released in January 2010 are listed. *(Source: Associated Press)*

1. Alabama 2. Texas 3. Florida 4. Boise State 5. Ohio State

20. Politics The three political parties in the 111th Congress are listed below.

Republican Democrat Independent

21. Top Salespeople The regions representing the top salespeople in a corporation for the past six years are given.

Southeast Northwest Northeast
Southeast Southwest Southwest

22. Fish Lengths The lengths (in inches) of a sample of striped bass caught in Maryland waters are listed. *(Adapted from National Marine Fisheries Service, Fisheries Statistics and Economics Division)*

16 17.25 19 18.75 21 20.3 19.8 24 21.82

23. Best Seller List The top five hardcover nonfiction books on *The New York Times* Best Seller List on January 19, 2010 are shown. *(Source: The New York Times)*

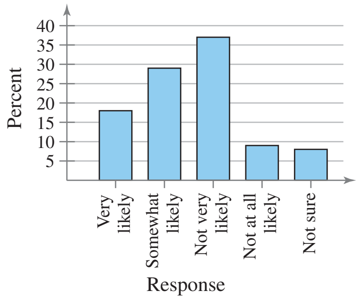
1. Committed 2. Have a Little Faith 3. The Checklist Manifesto
4. Going Rogue 5. Stones Into Schools

24. Ticket Prices The average ticket prices for 10 Broadway shows in 2009 are listed. *(Adapted from The Broadway League)*

$149 $128 $124 $91 $96 $106 $112 $95 $86 $74

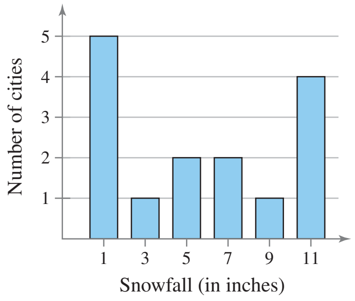
Graphical Analysis *In Exercises 25–28, identify the level of measurement of the data listed on the horizontal axis in the graph.*

25. **Over the Next Few Years, How Likely Is It That the United States Will Enter a 1930s-Like Depression?**



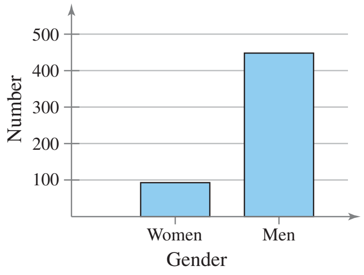
(Source: Rasmussen Reports)

26. **Average January Snowfall for 15 Cities**



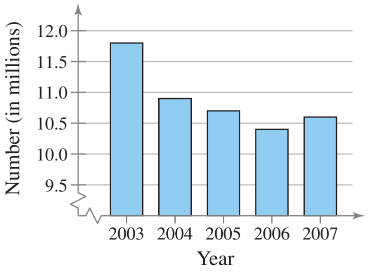
(Source: National Climatic Data Center)

27. **Gender Profile of the 111th Congress**



(Source: Congressional Research Service)

28. **Motor Vehicle Accidents by Year**



(Source: National Safety Council)

29. The following items appear on a physician's intake form. Identify the level of measurement of the data.

a. Temperature **b.** Allergies
c. Weight **d.** Pain level (scale of 0 to 10)

30. The following items appear on an employment application. Identify the level of measurement of the data.

a. Highest grade level completed **b.** Gender
c. Year of college graduation **d.** Number of years at last job

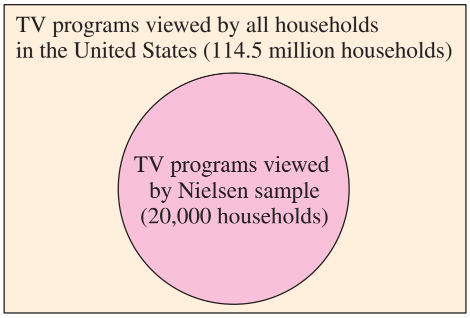
EXTENDING CONCEPTS

31. Writing What is an inherent zero? Describe three examples of data sets that have inherent zeros and three that do not.

32. Writing Describe two examples of data sets for each of the four levels of measurement. Justify your answer.

Rating Television Shows in the United States

The Nielsen Company has been rating television programs for more than 60 years. Nielsen uses several sampling procedures, but its main one is to track the viewing patterns of 20,000 households. These contain more than 45,000 people and are chosen to form a cross section of the overall population. The households represent various locations, ethnic groups, and income brackets. The data gathered from the Nielsen sample of 20,000 households are used to draw inferences about the population of all households in the United States.


TV programs viewed by all households in the United States (114.5 million households)
TV programs viewed by Nielsen sample (20,000 households)

Top-Ranked Programs in Overall Viewing for the Week of 11/23/09–11/29/09

Rank	Rank Last Week	Program Name	Network	Day, Time	Rating	Share	Audience
1	2	Dancing with the Stars	ABC	Mon., 8:00 P.M.	12.9	19	20,411,000
2	1	NCIS	CBS	Tues., 8:00 P.M.	12.3	20	20,348,000
3	4	Dancing with the Stars Results	ABC	Tues., 9:00 P.M.	12.0	20	19,294,000
4	3	NBC Sunday Night Football	NBC	Sun., 8:15 P.M.	11.5	18	19,210,000
5	8	NCIS: Los Angeles	CBS	Tues., 9:00 P.M.	10.4	16	17,221,000
6	5	60 Minutes	CBS	Sun., 7:00 P.M.	9.0	14	14,377,000
7	15	The Big Bang Theory	CBS	Mon., 9:30 P.M.	8.4	13	14,129,000
8	16	Sunday Night NFL Pre-Kick	NBC	Sun., 8:00 P.M.	8.4	13	13,927,000
9	12	Two and a Half Men	CBS	Mon., 9:00 P.M.	8.3	12	13,877,000
10	11	Criminal Minds	CBS	Wed., 9:00 P.M.	8.2	14	13,605,000

Copyrighted information of The Nielsen Company, licensed for use herein.

EXERCISES

1. **Rating Points** Each rating point represents 1,145,000 households, or 1% of the households in the United States. Does a program with a rating of 8.4 have twice the number of households as a program with a rating of 4.2? Explain your reasoning.

2. **Sampling Percent** What percentage of the total number of U.S. households is used in the Nielsen sample?

3. **Nominal Level of Measurement** Which columns in the table contain data at the nominal level?

4. **Ordinal Level of Measurement** Which columns in the table contain data at the ordinal level? Describe two ways that the data can be ordered.

5. **Interval Level of Measurement** Which column in the table contains data at the interval level? How can these data be ordered?

6. **Ratio Level of Measurement** Which columns contain data at the ratio level?

7. **Rankings** The column listed as "Share" gives the percentage of televisions in use at a given time. The 11th ranked program for this week is CSI: Miami with a rating of 8.4 and share of 14. Using this information, how does Nielsen rank the programs? Why do you think they do it this way? Explain your reasoning.

8. **Inferences** What decisions (inferences) can be made on the basis of the Nielsen ratings?

1.3 Data Collection and Experimental Design

WHAT YOU SHOULD LEARN

- How to design a statistical study
- How to collect data by doing an observational study, performing an experiment, using a simulation, or using a survey
- How to design an experiment
- How to create a sample using random sampling, simple random sampling, stratified sampling, cluster sampling, and systematic sampling and how to identify a biased sample

Design of a Statistical Study ▸ Data Collection ▸ Experimental Design ▸ Sampling Techniques

DESIGN OF A STATISTICAL STUDY

The goal of every statistical study is to collect data and then use the data to make a decision. Any decision you make using the results of a statistical study is only as good as the process used to obtain the data. If the process is flawed, then the resulting decision is questionable.

Although you may never have to develop a statistical study, it is likely that you will have to interpret the results of one. And before you interpret the results of a study, you should determine whether the results are valid, as well as reliable. In other words, you should be familiar with how to design a statistical study.

GUIDELINES

Designing a Statistical Study

1. Identify the variable(s) of interest (the focus) and the population of the study.
2. Develop a detailed plan for collecting data. If you use a sample, make sure the sample is representative of the population.
3. Collect the data.
4. Describe the data, using descriptive statistics techniques.
5. Interpret the data and make decisions about the population using inferential statistics.
6. Identify any possible errors.

DATA COLLECTION

There are several ways you can collect data. Often, the focus of the study dictates the best way to collect data. The following is a brief summary of four methods of data collection.

- ***Do an observational study*** In an **observational study,** a researcher observes and measures characteristics of interest of part of a population but does not change existing conditions. For instance, an observational study was performed in which researchers observed and recorded the mouthing behavior on nonfood objects of children up to three years old. *(Source: Pediatrics Magazine)*
- ***Perform an experiment*** In performing an **experiment,** a **treatment** is applied to part of a population and responses are observed. Another part of the population may be used as a **control group,** in which no treatment is applied. In many cases, subjects (sometimes called **experimental units**) in the control group are given a **placebo,** which is a harmless, unmedicated treatment, that is made to look like the real treatment. The responses of the treatment group and control group can then be compared and studied. In most cases, it is a good idea to use the same number of subjects for each treatment. For instance, an experiment was performed in which diabetics took cinnamon extract daily while a control group took none. After 40 days, the diabetics who took the cinnamon reduced their risk of heart disease while the control group experienced no change. *(Source: Diabetes Care)*

INSIGHT

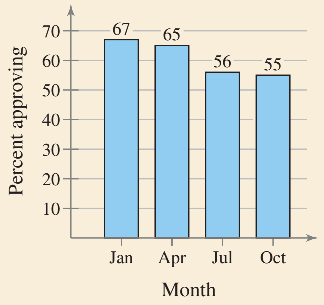
In an observational study, a researcher does not influence the responses. In an experiment, a researcher deliberately applies a treatment before observing the responses.

- ***Use a simulation*** A **simulation** is the use of a mathematical or physical model to reproduce the conditions of a situation or process. Collecting data often involves the use of computers. Simulations allow you to study situations that are impractical or even dangerous to create in real life, and often they save time and money. For instance, automobile manufacturers use simulations with dummies to study the effects of crashes on humans. Throughout this course, you will have the opportunity to use applets that simulate statistical processes on a computer.
- ***Use a survey*** A **survey** is an investigation of one or more characteristics of a population. Most often, surveys are carried out on *people* by asking them questions. The most common types of surveys are done by interview, mail, or telephone. In designing a survey, it is important to word the questions so that they do not lead to biased results, which are not representative of a population. For instance, a survey is conducted on a sample of female physicians to determine whether the primary reason for their career choice is financial stability. In designing the survey, it would be acceptable to make a list of reasons and ask each individual in the sample to select her first choice.

PICTURING THE WORLD

The Gallup Organization conducts many polls (or surveys) regarding the president, Congress, and political and nonpolitical issues. A commonly cited Gallup poll is the public approval rating of the president. For instance, the approval ratings for President Barack Obama throughout 2009 are shown in the following graph. (The rating is from the poll conducted at the end of each month.)

President's Approval Ratings, 2009



Discuss some ways that Gallup could select a biased sample to conduct a poll. How could Gallup select a sample that is unbiased?

EXAMPLE 1

Deciding on Methods of Data Collection

Consider the following statistical studies. Which method of data collection would you use to collect data for each study? Explain your reasoning.

1. A study of the effect of changing flight patterns on the number of airplane accidents
2. A study of the effect of eating oatmeal on lowering blood pressure
3. A study of how fourth grade students solve a puzzle
4. A study of U.S. residents' approval rating of the U.S. president

Solution

1. Because it is impractical to create this situation, use a simulation.
2. In this study, you want to measure the effect a treatment (eating oatmeal) has on patients. So, you would want to perform an experiment.
3. Because you want to observe and measure certain characteristics of part of a population, you could do an observational study.
4. You could use a survey that asks, "Do you approve of the way the president is handling his job?"

Try It Yourself 1

Consider the following statistical studies. Which method of data collection would you use to collect data for each study?

1. A study of the effect of exercise on relieving depression
2. A study of the success of graduates of a large university in finding a job within one year of graduation

a. Identify the *focus* of the study.
b. Identify the *population* of the study.
c. Choose an appropriate *method of data collection*.

Answer: Page A30

EXPERIMENTAL DESIGN

In order to produce meaningful unbiased results, experiments should be carefully designed and executed. It is important to know what steps should be taken to make the results of an experiment valid. Three key elements of a well-designed experiment are *control*, *randomization*, and *replication*.

Because experimental results can be ruined by a variety of factors, being able to control these influential factors is important. One such factor is a *confounding variable*.

DEFINITION

A **confounding variable** occurs when an experimenter cannot tell the difference between the effects of different factors on a variable.

INSIGHT

The **Hawthorne effect** occurs in an experiment when subjects change their behavior simply because they know they are participating in an experiment.

For instance, to attract more customers, a coffee shop owner experiments by remodeling her shop using bright colors. At the same time, a shopping mall nearby has its grand opening. If business at the coffee shop increases, it cannot be determined whether it is because of the new colors or the new shopping mall. The effects of the colors and the shopping mall have been confounded.

Another factor that can affect experimental results is the *placebo effect*. The **placebo effect** occurs when a subject reacts favorably to a placebo when in fact the subject has been given no medicated treatment at all. To help control or minimize the placebo effect, a technique called *blinding* can be used.

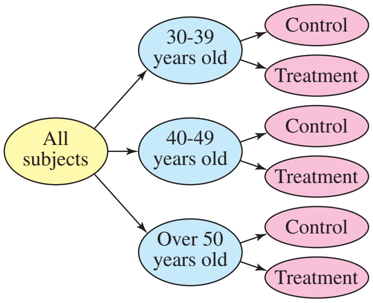
DEFINITION

Blinding is a technique where the subjects do not know whether they are receiving a treatment or a placebo. In a **double-blind experiment,** neither the experimenter nor the subjects know if the subjects are receiving a treatment or a placebo. The experimenter is informed after all the data have been collected. This type of experimental design is preferred by researchers.

Another technique that can be used to obtain unbiased results is *randomization*.

DEFINITION

Randomization is a process of randomly assigning subjects to different treatment groups.



Randomized Block Design

In a **completely randomized design,** subjects are assigned to different treatment groups through random selection. In some experiments, it may be necessary for the experimenter to use **blocks,** which are groups of subjects with similar characteristics. A commonly used experimental design is a **randomized block design.** To use a randomized block design, you should divide subjects with similar characteristics into blocks, and then, within each block, randomly assign subjects to treatment groups. For instance, an experimenter who is testing the effects of a new weight loss drink may first divide the subjects into age categories such as 30–39 years old, 40–49 years old, and over 50 years old, and then, within each age group, randomly assign subjects to either the treatment group or the control group as shown.

Another type of experimental design is a **matched-pairs design,** where subjects are paired up according to a similarity. One subject in the pair is randomly selected to receive one treatment while the other subject receives a different treatment. For instance, two subjects may be paired up because of their age, geographical location, or a particular physical characteristic.

Sample size, which is the number of subjects, is another important part of experimental design. To improve the validity of experimental results, *replication* is required.

INSIGHT

The validity of an experiment refers to the accuracy and reliability of the experimental results. The results of a valid experiment are more likely to be accepted in the scientific community.

DEFINITION

Replication is the repetition of an experiment under the same or similar conditions.

For instance, suppose an experiment is designed to test a vaccine against a strain of influenza. In the experiment, 10,000 people are given the vaccine and another 10,000 people are given a placebo. Because of the sample size, the effectiveness of the vaccine would most likely be observed. But, if the subjects in the experiment are not selected so that the two groups are similar (according to age and gender), the results are of less value.

EXAMPLE 2

▸ Analyzing an Experimental Design

A company wants to test the effectiveness of a new gum developed to help people quit smoking. Identify a potential problem with the given experimental design and suggest a way to improve it.

1. The company identifies ten adults who are heavy smokers. Five of the subjects are given the new gum and the other five subjects are given a placebo. After two months, the subjects are evaluated and it is found that the five subjects using the new gum have quit smoking.
2. The company identifies one thousand adults who are heavy smokers. The subjects are divided into blocks according to gender. Females are given the new gum and males are given the placebo. After two months, a significant number of the female subjects have quit smoking.

▸ Solution

1. The sample size being used is not large enough to validate the results of the experiment. The experiment must be replicated to improve the validity.
2. The groups are not similar. The new gum may have a greater effect on women than on men, or vice versa. The subjects can be divided into blocks according to gender, but then, within each block, they must be randomly assigned to be in the treatment group or in the control group.

▸ Try It Yourself 2

Using the information in Example 2, suppose the company identifies 240 adults who are heavy smokers. The subjects are randomly assigned to be in a treatment group or in a control group. Each subject is also given a DVD featuring the dangers of smoking. After four months, most of the subjects in the treatment group have quit smoking.

a. Identify a *potential problem* with the experimental design.
b. How could the design be *improved*? *Answer: Page A30*

SAMPLING TECHNIQUES

A **census** is a count or measure of an *entire* population. Taking a census provides complete information, but it is often costly and difficult to perform. A **sampling** is a count or measure of *part* of a population, and is more commonly used in statistical studies. To collect unbiased data, a researcher must ensure that the sample is representative of the population. Appropriate sampling techniques must be used to ensure that inferences about the population are valid. Remember that when a study is done with faulty data, the results are questionable. Even with the best methods of sampling, a **sampling error** may occur. A sampling error is the difference between the results of a sample and those of the population. When you learn about inferential statistics, you will learn techniques of controlling sampling errors.

A **random sample** is one in which every member of the population has an equal chance of being selected. A **simple random sample** is a sample in which every possible sample of the same size has the same chance of being selected. One way to collect a simple random sample is to assign a different number to each member of the population and then use a random number table like the one in Appendix B. Responses, counts, or measures for members of the population whose numbers correspond to those generated using the table would be in the sample. Calculators and computer software programs are also used to generate random numbers (see page 34).

INSIGHT

A **biased sample** is one that is not representative of the population from which it is drawn. For instance, a sample consisting of only 18- to 22-year-old college students would not be representative of the entire 18- to 22-year-old population in the country.

To explore this topic further, see Activity 1.3 on page 26.

Table 1—Random Numbers

92630	78240	19267	95457	53497	23894	37708	79862
79445	78735	71549	44843	26104	67318	00701	34986
59654	71966	27386	50004	05358	94031	29281	18544
31524	49587	76612	39789	13537	48086	59483	60680
06348	76938	90379	51392	55887	71015	09209	79157

Portion of Table 1 found in Appendix B

Consider a study of the number of people who live in West Ridge County. To use a simple random sample to count the number of people who live in West Ridge County households, you could assign a different number to each household, use a technology tool or table of random numbers to generate a sample of numbers, and then count the number of people living in each selected household.

STUDY TIP

Here are instructions for using the random integer generator on a TI-83/84 Plus for Example 3.

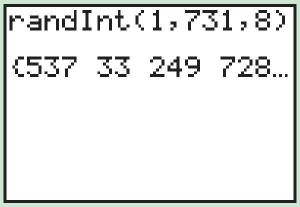
MATH

Choose the PRB menu.

5: randInt(

1 , 7 3 1 , 8)

ENTER



```
randInt(1,731,8)
{537 33 249 728…
```

Continuing to press ENTER will generate more random samples of 8 integers.

EXAMPLE 3

SC Report 1

Using a Simple Random Sample

There are 731 students currently enrolled in a statistics course at your school. You wish to form a sample of eight students to answer some survey questions. Select the students who will belong to the simple random sample.

Solution

Assign numbers 1 to 731 to the students in the course. In the table of random numbers, choose a starting place at random and read the digits in groups of three (because 731 is a three-digit number). For instance, if you started in the third row of the table at the beginning of the second column, you would group the numbers as follows:

719|66 2|738|6 50|004| 053|58 9|403|1 29|281| 185|44

Ignoring numbers greater than 731, the first eight numbers are 719, 662, 650, 4, 53, 589, 403, and 129. The students assigned these numbers will make up the sample. To find the sample using a TI-83/84 Plus, follow the instructions in the margin.

▸ Try It Yourself 3

A company employs 79 people. Choose a simple random sample of five to survey.

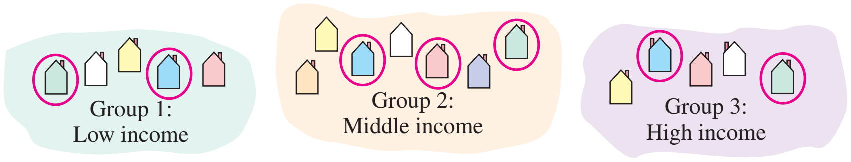
a. In the table in Appendix B, randomly choose a *starting place*.
b. *Read the digits* in groups of two.
c. Write the five random numbers.

Answer: Page A30

When you choose members of a sample, you should decide whether it is acceptable to have the same population member selected more than once. If it is acceptable, then the sampling process is said to be *with replacement*. If it is not acceptable, then the sampling process is said to be *without replacement*.

There are several other commonly used sampling techniques. Each has advantages and disadvantages.

- ***Stratified Sample*** When it is important for the sample to have members from each segment of the population, you should use a stratified sample. Depending on the focus of the study, members of the population are divided into two or more subsets, called *strata*, that share a similar characteristic such as age, gender, ethnicity, or even political preference. A sample is then randomly selected from each of the strata. Using a stratified sample ensures that each segment of the population is represented. For instance, to collect a stratified sample of the number of people who live in West Ridge County households, you could divide the households into socioeconomic levels, and then randomly select households from each level.



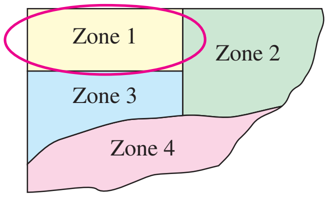
Stratified Sampling

- ***Cluster Sample*** When the population falls into naturally occurring subgroups, each having similar characteristics, a cluster sample may be the most appropriate. To select a cluster sample, divide the population into groups, called *clusters*, and select all of the members in one or more (but not all) of the clusters. Examples of clusters could be different sections of the same course or different branches of a bank. For instance, to collect a cluster sample of the number of people who live in West Ridge County households, divide the households into groups according to zip codes, then select all the households in one or more, but not all, zip codes and count the number of people living in each household. In using a cluster sample, care must be taken to ensure that all clusters have similar characteristics. For instance, if one of the zip code clusters has a greater proportion of high-income people, the data might not be representative of the population.

INSIGHT

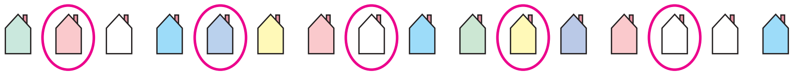
For stratified sampling, each of the strata contains members with a certain characteristic (for instance, a particular age group). In contrast, clusters consist of geographic groupings, and each cluster should contain members with all of the characteristics (for instance, all age groups). With stratified samples, some of the members of each group are used. In a cluster sampling, all of the members of one or more groups are used.

Zip Code Zones in West Ridge County



Cluster Sampling

- ***Systematic Sample*** A systematic sample is a sample in which each member of the population is assigned a number. The members of the population are ordered in some way, a starting number is randomly selected, and then sample members are selected at regular intervals from the starting number. (For instance, every 3rd, 5th, or 100th member is selected.) For instance, to collect a systematic sample of the number of people who live in West Ridge County households, you could assign a different number to each household, randomly choose a starting number, select every 100th household, and count the number of people living in each. An advantage of systematic sampling is that it is easy to use. In the case of any regularly occurring pattern in the data, however, this type of sampling should be avoided.



Systematic Sampling

A type of sample that often leads to biased studies (so it is not recommended) is a **convenience sample.** A convenience sample consists only of available members of the population.

EXAMPLE 4

Identifying Sampling Techniques

You are doing a study to determine the opinions of students at your school regarding stem cell research. Identify the sampling technique you are using if you select the samples listed. Discuss potential sources of bias (if any). Explain.

1. You divide the student population with respect to majors and randomly select and question some students in each major.

2. You assign each student a number and generate random numbers. You then question each student whose number is randomly selected.

3. You select students who are in your biology class.

Solution

1. Because students are divided into strata (majors) and a sample is selected from each major, this is a stratified sample.

2. Each sample of the same size has an equal chance of being selected and each student has an equal chance of being selected, so this is a simple random sample.

3. Because the sample is taken from students that are readily available, this is a convenience sample. The sample may be biased because biology students may be more familiar with stem cell research than other students and may have stronger opinions.

Try It Yourself 4

You want to determine the opinions of students regarding stem cell research. Identify the sampling technique you are using if you select the samples listed.

1. You select a class at random and question each student in the class.

2. You assign each student a number and, after choosing a starting number, question every 25th student.

a. Determine *how* the sample is *selected* and identify the corresponding *sampling technique*.
b. Discuss potential sources of *bias* (if any). Explain. *Answer: Page A30*

1.3 EXERCISES

BUILDING BASIC SKILLS AND VOCABULARY

1. What is the difference between an observational study and an experiment?
2. What is the difference between a census and a sampling?
3. What is the difference between a random sample and a simple random sample?
4. What is replication in an experiment, and why is it important?

True or False? *In Exercises 5–10, determine whether the statement is true or false. If it is false, rewrite it as a true statement.*

5. In a randomized block design, subjects with similar characteristics are divided into blocks, and then, within each block, randomly assigned to treatment groups.
6. A double-blind experiment is used to increase the placebo effect.
7. Using a systematic sample guarantees that members of each group within a population will be sampled.
8. A census is a count of part of a population.
9. The method for selecting a stratified sample is to order a population in some way and then select members of the population at regular intervals.
10. To select a cluster sample, divide a population into groups and then select all of the members in at least one (but not all) of the groups.

Deciding on the Method of Data Collection *In Exercises 11–16, explain which method of data collection you would use to collect data for the study.*

11. A study of the health of 168 kidney transplant patients at a hospital
12. A study of motorcycle helmet usage in a city without a helmet law
13. A study of the effect on the human digestive system of potato chips made with a fat substitute
14. A study of the effect of a product's warning label to determine whether consumers will still buy the product
15. A study of how fast a virus would spread in a metropolitan area
16. A study of how often people wash their hands in public restrooms

USING AND INTERPRETING CONCEPTS

17. **Allergy Drug** A pharmaceutical company wants to test the effectiveness of a new allergy drug. The company identifies 250 females 30–35 years old who suffer from severe allergies. The subjects are randomly assigned into two groups. One group is given the new allergy drug and the other is given a placebo that looks exactly like the new allergy drug. After six months, the subjects' symptoms are studied and compared.
    (a) Identify the experimental units and treatments used in this experiment.
    (b) Identify a potential problem with the experimental design being used and suggest a way to improve it.
    (c) How could this experiment be designed to be double-blind?

18. **Sneakers** Nike developed a new type of sneaker designed to help delay the onset of arthritis in the knee. Eighty people with early signs of arthritis volunteered for a study. One-half of the volunteers wore the experimental sneakers and the other half wore regular Nike sneakers that looked exactly like the experimental sneakers. The individuals wore the sneakers every day. At the conclusion of the study, their symptoms were evaluated and MRI tests were performed on their knees. *(Source: Washington Post)*

(a) Identify the experimental units and treatments used in this experiment.

(b) Identify a potential problem with the experimental design being used and suggest a way to improve it.

(c) The experiment is described as a placebo-controlled, double-blind study. Explain what this means.

(d) Of the 80 volunteers, suppose 40 are men and 40 are women. How could blocking be used in designing this experiment?

Identifying Sampling Techniques *In Exercises 19–26, identify the sampling technique used, and discuss potential sources of bias (if any). Explain.*

19. Using random digit dialing, researchers call 1400 people and ask what obstacles (such as childcare) keep them from exercising.

20. Chosen at random, 500 rural and 500 urban persons age 65 or older are asked about their health and their experience with prescription drugs.

21. Questioning students as they leave a university library, a researcher asks 358 students about their drinking habits.

22. After a hurricane, a disaster area is divided into 200 equal grids. Thirty of the grids are selected, and every occupied household in the grid is interviewed to help focus relief efforts on what residents require the most.

23. Chosen at random, 580 customers at a car dealership are contacted and asked their opinions of the service they received.

24. Every tenth person entering a mall is asked to name his or her favorite store.

25. Soybeans are planted on a 48-acre field. The field is divided into one-acre subplots. A sample is taken from each subplot to estimate the harvest.

26. From calls made with randomly generated telephone numbers, 1012 respondents are asked if they rent or own their residences.

27. **Random Number Table** Use the seventh row of Table 1 in Appendix B to generate 12 random numbers between 1 and 99.

28. **Random Number Table** Use the twelfth row of Table 1 in Appendix B to generate 10 random numbers between 1 and 920.

29. **Sleep Deprivation** A researcher wants to study the effects of sleep deprivation on motor skills. Eighteen people volunteer for the experiment: Jake, Maria, Mike, Lucy, Ron, Adam, Bridget, Carlos, Steve, Susan, Vanessa, Rick, Dan, Kate, Pete, Judy, Mary, and Connie. Use a random number generator to choose nine subjects for the treatment group. The other nine subjects will go into the control group. List the subjects in each group. Tell which method you would use to generate the random numbers.

30. **Random Number Generation** Volunteers for an experiment are numbered from 1 to 70. The volunteers are to be randomly assigned to two different treatment groups. Use a random number generator different from the one you used in Exercise 29 to choose 35 subjects for the treatment group. The other 35 subjects will go into the control group. List the subjects, according to number, in each group. Tell which method you used to generate the random numbers.

Choosing Between a Census and a Sampling *In Exercises 31 and 32, determine whether you would take a census or use a sampling. If you would use a sampling, decide what sampling technique you would use. Explain your reasoning.*

31. The average age of the 115 residents of a retirement community

32. The most popular type of movie among 100,000 online movie rental subscribers

Recognizing a Biased Question *In Exercises 33–36, determine whether the survey question is biased. If the question is biased, suggest a better wording.*

33. Why does eating whole-grain foods improve your health?

34. Why does text messaging while driving increase the risk of a crash?

35. How much do you exercise during an average week?

36. Why do you think the media have a negative effect on teen girls' dieting habits?

37. Writing A sample of television program ratings by The Nielsen Company is described on page 15. Discuss the strata used in the sample. Why is it important to have a stratified sample for these ratings?

SC **38.** Use StatCrunch to generate the following random numbers.

a. 8 numbers between 1 and 50

b. 15 numbers between 1 and 150

c. 16 numbers between 1 and 325

d. 20 numbers between 1 and 1000

EXTENDING CONCEPTS

39. Observational studies are sometimes referred to as *natural experiments*. Explain, in your own words, what this means.

40. Open and Closed Questions Two types of survey questions are open questions and closed questions. An open question allows for any kind of response; a closed question allows for only a fixed response. An open question, and a closed question with its possible choices, are given below. List an advantage and a disadvantage of each question.

Open Question What can be done to get students to eat healthier foods?

Closed Question How would you get students to eat healthier foods?

1. Mandatory nutrition course
2. Offer only healthy foods in the cafeteria and remove unhealthy foods
3. Offer more healthy foods in the cafeteria and raise the prices on unhealthy foods

41. Who Picked These People? Some polling agencies ask people to call a telephone number and give their response to a question. (a) List an advantage and a disadvantage of a survey conducted in this manner. (b) What sampling technique is used in such a survey?

42. Give an example of an experiment where confounding may occur.

43. Why is it important to use blinding in an experiment?

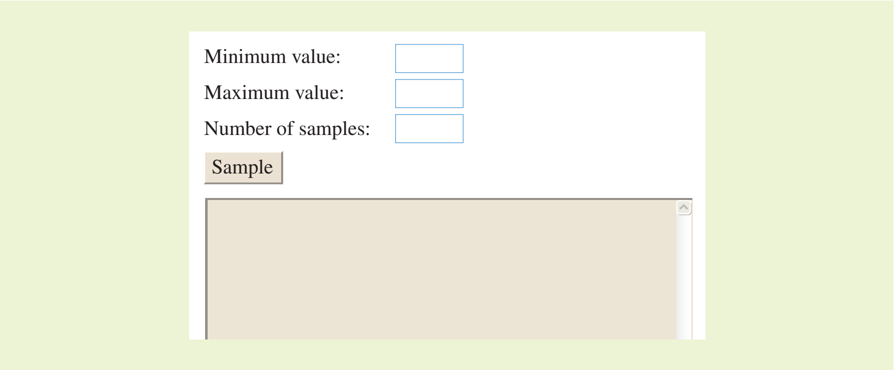
44. How are the placebo effect and the Hawthorne effect similar? How are they different?

45. How is a randomized block design in experiments similar to a stratified sample?

ACTIVITY 1.3 Random Numbers

APPLET

The *random numbers* applet is designed to allow you to generate random numbers from a range of values. You can specify integer values for the minimum value, maximum value, and the number of samples in the appropriate fields. You should not use decimal points when filling in the fields. When SAMPLE is clicked, the applet generates random values, which are displayed as a list in the text field.



Minimum value:

Maximum value:

Number of samples:

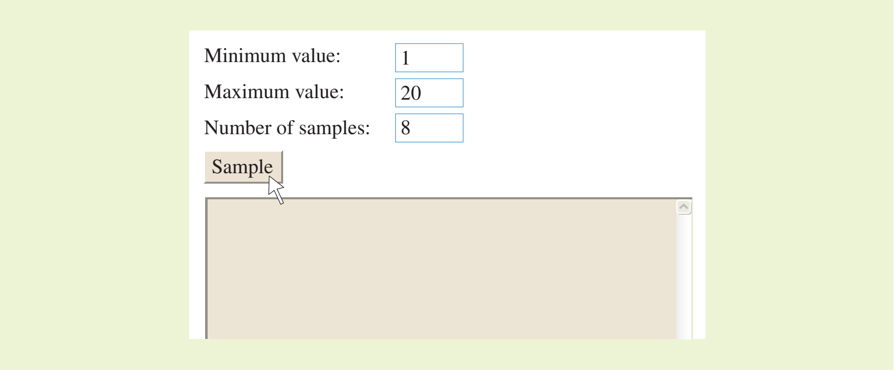
Explore

Step 1 Specify a minimum value.
Step 2 Specify a maximum value.
Step 3 Specify the number of samples.
Step 4 Click SAMPLE to generate a list of random values.

Draw Conclusions

APPLET

1. Specify the minimum, maximum, and number of samples to be 1, 20, and 8, respectively, as shown. Run the applet. Continue generating lists until you obtain one that shows that the random sample is taken with replacement. Write down this list. How do you know that the list is a random sample taken with replacement?



Minimum value: 1

Maximum value: 20

Number of samples: 8

2. Use the applet to repeat Example 3 on page 20. What values did you use for the minimum, maximum, and number of samples? Which method do you prefer? Explain.

USES AND ABUSES

Uses

Experiments with Favorable Results An experiment that began in March 2003 studied 321 women with advanced breast cancer. All of the women had been previously treated with other drugs, but the cancer had stopped responding to the medications. The women were then given the opportunity to take a new drug combined with a particular chemotherapy drug.

The subjects were divided into two groups, one that took the new drug combined with a chemotherapy drug, and one that took only the chemotherapy drug. After three years, results showed that the new drug in combination with the chemotherapy drug delayed the progression of cancer in the subjects. The results were so significant that the study was stopped, and the new drug was offered to all women in the study. The Food and Drug Administration has since approved use of the new drug in conjunction with a chemotherapy drug.

Abuses

Experiments with Unfavorable Results From 1988 to 1991, one hundred eighty thousand teenagers in Norway were used as subjects to test a new vaccine against the deadly bacteria *meningococcus b*. A brochure describing the possible effects of the vaccine stated, "it is unlikely to expect serious complications," while information provided to the Norwegian Parliament stated, "serious side effects can not be excluded." The vaccine trial had some disastrous results: More than 500 side effects were reported, with some considered serious, and several of the subjects developed serious neurological diseases. The results showed that the vaccine was providing immunity in only 57% of the cases. This result was not sufficient for the vaccine to be added to Norway's vaccination program. Compensations have since been paid to the vaccine victims.

Ethics

Experiments help us further understand the world that surrounds us. But, in some cases, they can do more harm than good. In the Norwegian experiments, several ethical questions arise. Was the Norwegian experiment unethical if the best interests of the subjects were neglected? When should the experiment have been stopped? Should it have been conducted at all? If serious side effects are not reported and are withheld from subjects, there is no ethical question here, it is just wrong.

On the other hand, the breast cancer researchers would not want to deny the new drug to a group of patients with a life-threatening disease. But again, questions arise. How long must a researcher continue an experiment that shows better-than-expected results? How soon can a researcher conclude a drug is safe for the subjects involved?

■ EXERCISES

1. ***Unfavorable Results*** Find an example of a real-life experiment that had unfavorable results. What could have been done to avoid the outcome of the experiment?

2. ***Stopping an Experiment*** In your opinion, what are some problems that may arise if clinical trials of a new experimental drug or vaccine are stopped early and then the drug or vaccine is distributed to other subjects or patients?

1 CHAPTER SUMMARY

What did you learn?	EXAMPLE(S)	REVIEW EXERCISES
Section 1.1		
■ How to distinguish between a population and a sample	*1*	*1–4*
■ How to distinguish between a parameter and a statistic	*2*	*5–8*
■ How to distinguish between descriptive statistics and inferential statistics	*3*	*9, 10*
Section 1.2		
■ How to distinguish between qualitative data and quantitative data	*1*	*11–16*
■ How to classify data with respect to the four levels of measurement: nominal, ordinal, interval, and ratio	*2, 3*	*17–20*
Section 1.3		
■ How data are collected: by doing an observational study, performing an experiment, using a simulation, or using a survey	*1*	*21–24*
■ How to design an experiment	*2*	*25, 26*
■ How to create a sample using random sampling, simple random sampling, stratified sampling, cluster sampling, and systematic sampling	*3, 4*	*27–34*
■ How to identify a biased sample	*4*	*35–38*

1 REVIEW EXERCISES

SECTION 1.1

In Exercises 1–4, identify the population and the sample.

1. A survey of 1000 U.S. adults found that 83% think credit cards tempt people to buy things they cannot afford. *(Source: Rasmussen Reports)*

2. Thirty-eight nurses working in the San Francisco area were surveyed concerning their opinions of managed health care.

3. A survey of 39 credit cards found that the average annual percentage rate (APR) is 12.83%. *(Source: Consumer Action)*

4. A survey of 1205 physicians found that about 60% had considered leaving the practice of medicine because they were discouraged over the state of U.S. health care. *(Source: The Physician Executive Journal of Medical Management)*

In Exercises 5–8, determine whether the numerical value describes a parameter or a statistic.

5. The 2009 team payroll of the Philadelphia Phillies was $113,004,046. *(Source: USA Today)*

6. In a survey of 752 adults in the United States, 42% think there should be a law that prohibits people from talking on cell phones in public places. *(Source: University of Michigan)*

7. In a recent study of math majors at a university, 10 students were minoring in physics.

8. Fifty percent of a sample of 1508 U.S. adults say they oppose drilling for oil and gas in the Arctic National Wildlife Refuge. *(Source: Pew Research Center)*

9. Which part of the study described in Exercise 3 represents the descriptive branch of statistics? Make an inference based on the results of the study.

10. Which part of the survey described in Exercise 4 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

SECTION 1.2

In Exercises 11–16, determine which data are qualitative data and which are quantitative data. Explain your reasoning.

11. The monthly salaries of the employees at an accounting firm

12. The Social Security numbers of the employees at an accounting firm

13. The ages of a sample of 350 employees of a software company

14. The zip codes of a sample of 350 customers at a sporting goods store

15. The 2010 revenues of the companies on the Fortune 500 list

16. The marital statuses of all professional golfers

In Exercises 17–20, identify the data set's level of measurement. Explain your reasoning.

17. The daily high temperatures (in degrees Fahrenheit) for Mohave, Arizona for a week in June are listed. *(Source: Arizona Meteorological Network)*

93 91 86 94 103 104 103

18. The levels of the Homeland Security Advisory System are listed.

Severe High Elevated Guarded Low

19. The four departments of a printing company are listed.

Administration   Sales   Production   Billing

20. The total compensations (in millions of dollars) of the top ten female CEOs in the United States are listed. *(Source: Forbes)*

9.4   5.3   11.8   11.1   9.4   4.1   6.6   5.7   4.6   4.5

SECTION 1.3

In Exercises 21–24, decide which method of data collection you would use to collect data for the study. Explain your reasoning.

21. A study of charitable donations of the CEOs in Syracuse, New York

22. A study of the effect of koalas on the ecosystem of Kangaroo Island, Australia

23. A study of how training dogs from animal shelters affects inmates at a prison

24. A study of college professors' opinions on teaching classes online

In Exercises 25 and 26, an experiment is being performed to test the effects of sleep deprivation on memory recall. Two hundred students volunteer for the experiment. The students will be placed in one of five different treatment groups, including the control group.

25. Explain how you could design an experiment so that it uses a randomized block design.

26. Explain how you could design an experiment so that it uses a completely randomized design.

27. Random Number Table Use the fifth row of Table 1 in Appendix B to generate 8 random numbers between 1 and 650.

28. Census or Sampling? You want to know the favorite spring break destination among 15,000 students at a university. Decide whether you would take a census or use a sampling. If you would use a sampling, decide what technique you would use. Explain your reasoning.

In Exercises 29–34, identify the sampling technique used in the study. Explain your reasoning.

29. Using random digit dialing, researchers ask 1003 U.S. adults their plans on working during retirement. *(Source: Princeton Survey Research Associates International)*

30. A student asks 18 friends to participate in a psychology experiment.

31. A pregnancy study in Cebu, Philippines randomly selects 33 communities from the Cebu metropolitan area, then interviews all available pregnant women in these communities. *(Adapted from Cebu Longitudinal Health and Nutrition Survey)*

32. Law enforcement officials stop and check the driver of every third vehicle for blood alcohol content.

33. Twenty-five students are randomly selected from each grade level at a high school and surveyed about their study habits.

34. A journalist interviews 154 people waiting at an airport baggage claim and asks them how safe they feel during air travel.

In Exercises 35–38, identify a bias or error that might occur in the indicated survey or study.

35. study in Exercise 29

36. experiment in Exercise 30

37. study in Exercise 31

38. sampling in Exercise 32

1 CHAPTER QUIZ

Take this quiz as you would take a quiz in class. After you are done, check your work against the answers given in the back of the book.

1. Identify the population and the sample in the following study.

A study of the dietary habits of 20,000 men was conducted to find a link between high intakes of dairy products and prostate cancer. *(Source: Harvard School of Public Health)*

2. Determine whether the numerical value is a parameter or a statistic.

(a) In a survey of 2253 Internet users, 19% use Twitter or another service to share social updates. *(Source: Pew Internet Project)*

(b) At a college, 90% of the Board of Trustees members approved the contract of the new president.

(c) A survey of 846 chief financial officers and senior comptrollers shows that 55% of U.S. companies are reducing bonuses. *(Source: Grant Thornton International)*

3. Determine whether the data are qualitative or quantitative.

(a) A list of debit card pin numbers

(b) The final scores on a video game

4. Identify each data set's level of measurement. Explain your reasoning.

(a) A list of badge numbers of police officers at a precinct

(b) The horsepowers of racing car engines

(c) The top 10 grossing films released in 2010

(d) The years of birth for the runners in the Boston marathon

5. Decide which method of data collection you would use to gather data for each study. Explain your reasoning.

(a) A study on the effect of low dietary intake of vitamin C and iron on lead levels in adults

(b) The ages of people living within 500 miles of your home

6. An experiment is being performed to test the effects of a new drug on high blood pressure. The experimenter identifies 320 people ages 35–50 years old with high blood pressure for participation in the experiment. The subjects are divided into equal groups according to age. Within each group, subjects are then randomly selected to be in either the treatment group or the control group. What type of experimental design is being used for this experiment?

7. Identify the sampling technique used in each study. Explain your reasoning.

(a) A journalist goes to a campground to ask people how they feel about air pollution.

(b) For quality assurance, every tenth machine part is selected from an assembly line and measured for accuracy.

(c) A study on attitudes about smoking is conducted at a college. The students are divided by class (freshman, sophomore, junior, and senior). Then a random sample is selected from each class and interviewed.

8. Which sampling technique used in Exercise 7 could lead to a biased study?

PUTTING IT ALL TOGETHER

Real Statistics — Real Decisions

You are a researcher for a professional research firm. Your firm has won a contract to do a study for an air travel industry publication. The editors of the publication would like to know their readers' thoughts on air travel factors such as ticket purchase, services, safety, comfort, economic growth, and security. They would also like to know the thoughts of adults who use air travel for business as well as for recreation.

The editors have given you their readership database and 20 questions they would like to ask (two sample questions from a previous study are given at the right). You know that it is too expensive to contact all of the readers, so you need to determine a way to contact a representative sample of the entire readership population.

How did you acquire your ticket?

Response	Percent
Travel agent	35.1%
Directly from airline	20.9%
Online, using the airline's website	21.0%
Online, from a travel site other than the airline	18.5%
Other	4.5%

(Source: Resource Systems Group)

How many associates, friends, or family members traveled together in your party?

Response	Percent
1 (traveled alone)	48.7%
2 (traveled with one other person)	29.7%
3 (traveled with 2 others)	7.1%
4 (traveled with 3 others)	7.7%
5 (traveled with 4 others)	3.0%
6 or more (traveled with 5 or more others)	3.8%

(Source: Resource Systems Group)

EXERCISES

1. *How Would You Do It?*

(a) What sampling technique would you use to select the sample for the study? Why?

(b) Will the technique you choose in part (a) give you a sample that is representative of the population?

(c) Describe the method for collecting data.

(d) Identify possible flaws or biases in your study.

2. *Data Classification*

(a) What type of data do you expect to collect: qualitative, quantitative, or both? Why?

(b) At what levels of measurement do you think the data in the study will be? Why?

(c) Will the data collected for the study represent a population or a sample?

(d) Will the numerical descriptions of the data be parameters or statistics?

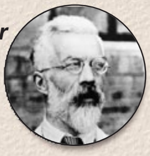
3. *How They Did It*

When the *Resource Systems Group* did a similar study, they used an Internet survey. They sent out 1000 invitations to participate in the survey and received 621 completed surveys.

(a) Describe some possible errors in collecting data by Internet surveys.

(b) Compare your method for collecting data in Exercise 1 to this method.

HISTORY OF STATISTICS - TIMELINE

CONTRIBUTOR	TIME	CONTRIBUTION
John Graunt *(1620–1674)*	17th century	Studied records of deaths in London in the early 1600s. The first to make extensive statistical observations from massive amounts of data (Chapter 2), his work laid the foundation for modern statistics.
Blaise Pascal *(1623–1662)* ***Pierre de Fermat*** *(1601–1665)*		Pascal and Fermat corresponded about basic probability problems (Chapter 3)—especially those dealing with gaming and gambling.
Pierre Laplace *(1749–1827)*	18th century	Studied probability (Chapter 3) and is credited with putting probability on a sure mathematical footing.
Carl Friedrich Gauss *(1777–1855)*		Studied regression and the method of least squares (Chapter 9) through astronomy. In his honor, the normal distribution is sometimes called the Gaussian distribution.
Lambert Quetelet *(1796–1874)*	19th century	Used descriptive statistics (Chapter 2) to analyze crime and mortality data and studied census techniques. Described normal distributions (Chapter 5) in connection with human traits such as height.
Francis Galton *(1822–1911)*		Used regression and correlation (Chapter 9) to study genetic variation in humans. He is credited with the discovery of the Central Limit Theorem (Chapter 5).
Karl Pearson *(1857–1936)*	20th century	Studied natural selection using correlation (Chapter 9). Formed first academic department of statistics and helped develop chi-square analysis (Chapter 6).
William Gosset *(1876–1937)*		Studied process of brewing and developed *t*-test to correct problems connected with small sample sizes (Chapter 6).
Charles Spearman *(1863–1945)*		British psychologist who was one of the first to develop intelligence testing using factor analysis (Chapter 10).
Ronald Fisher *(1890–1962)* 		Studied biology and natural selection and developed ANOVA (Chapter 10), stressed the importance of experimental design (Chapter 1), and was the first to identify the null and alternative hypotheses (Chapter 7).
Frank Wilcoxon *(1892–1965)*	20th century (later)	Biochemist who used statistics to study plant pathology. He introduced two-sample tests (Chapter 8), which led the way to the development of nonparametric statistics.
John Tukey *(1915–2000)*		Worked at Princeton during World War II. Introduced exploratory data analysis techniques such as stem-and-leaf plots (Chapter 2). Also, worked at Bell Laboratories and is best known for his work in inferential statistics (Chapters 6–11).
David Kendall *(1918–2007)*		Worked at Princeton and Cambridge. Was a leading authority on applied probability and data analysis (Chapters 2 and 3).

TECHNOLOGY

USING TECHNOLOGY IN STATISTICS

With large data sets, you will find that calculators or computer software programs can help perform calculations and create graphics. Of the many calculators and statistical software programs that are available, we have chosen to incorporate the TI-83/84 Plus graphing calculator, and MINITAB and Excel software into this text.

The following example shows how to use these three technologies to generate a list of random numbers. This list of random numbers can be used to select sample members or perform simulations.

EXAMPLE

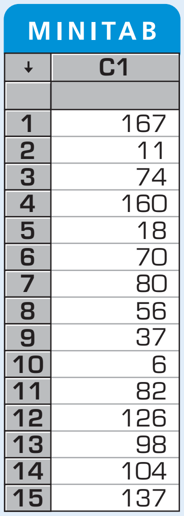
Generating a List of Random Numbers

A quality control department inspects a random sample of 15 of the 167 cars that are assembled at an auto plant. How should the cars be chosen?

Solution

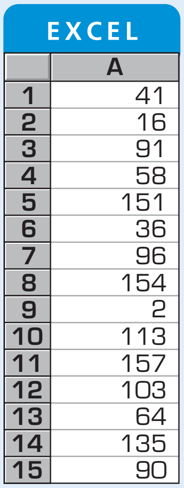
One way to choose the sample is to first number the cars from 1 to 167. Then you can use technology to form a list of random numbers from 1 to 167. Each of the technology tools shown requires different steps to generate the list. Each, however, does require that you identify the minimum value as 1 and the maximum value as 167. Check your user's manual for specific instructions.

MINITAB



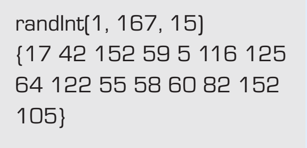
↓	C1
1	167
2	11
3	74
4	160
5	18
6	70
7	80
8	56
9	37
10	6
11	82
12	126
13	98
14	104
15	137

EXCEL



	A
1	41
2	16
3	91
4	58
5	151
6	36
7	96
8	154
9	2
10	113
11	157
12	103
13	64
14	135
15	90

TI-83/84 PLUS



```
randInt(1, 167, 15)
{17 42 152 59 5 116 125
64 122 55 58 60 82 152
105}
```

Recall that when you generate a list of random numbers, you should decide whether it is acceptable to have numbers that repeat. If it is acceptable, then the sampling process is said to be with replacement. If it is not acceptable, then the sampling process is said to be without replacement.

With each of the three technology tools shown on page 34, you have the capability of sorting the list so that the numbers appear in order. Sorting helps you see whether any of the numbers in the list repeat. If it is not acceptable to have repeats, you should specify that the tool generate more random numbers than you need.

EXERCISES

1. The SEC (Securities and Exchange Commission) is investigating a financial services company. The company being investigated has 86 brokers. The SEC decides to review the records for a random sample of 10 brokers. Describe how this investigation could be done. Then use technology to generate a list of 10 random numbers from 1 to 86 and order the list.

2. A quality control department is testing 25 smartphones from a shipment of 300 smartphones. Describe how this test could be done. Then use technology to generate a list of 25 random numbers from 1 to 300 and order the list.

3. Consider the population of ten digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. Select three random samples of five digits from this list. Find the average of each sample. Compare your results with the average of the entire population. Comment on your results. (*Hint:* To find the average, sum the data entries and divide the sum by the number of entries.)

4. Consider the population of 41 whole numbers from 0 to 40. What is the average of these numbers? Select three random samples of seven numbers from this list. Find the average of each sample. Compare your results with the average of the entire population. Comment on your results. (*Hint:* To find the average, sum the data entries and divide the sum by the number of entries.)

5. Use random numbers to simulate rolling a six-sided die 60 times. How many times did you obtain each number from 1 to 6? Are the results what you expected?

6. You rolled a six-sided die 60 times and got the following tally.

20 ones	20 twos	15 threes
3 fours	2 fives	0 sixes

Does this seem like a reasonable result? What inference might you draw from the result?

7. Use random numbers to simulate tossing a coin 100 times. Let 0 represent heads, and let 1 represent tails. How many times did you obtain each number? Are the results what you expected?

8. You tossed a coin 100 times and got 77 heads and 23 tails. Does this seem like a reasonable result? What inference might you draw from the result?

9. A political analyst would like to survey a sample of the registered voters in a county. The county has 47 election districts. How could the analyst use random numbers to obtain a cluster sample?

Extended solutions are given in the *Technology Supplement.*
Technical instruction is provided for MINITAB, Excel, and the TI-83/84 Plus.