Descriptive Statistics

CHAPTER 1

Introduction to Statistics

Clark County, Nevada has one of the fastest growing populations in the United States. Las Vegas is located in this county.

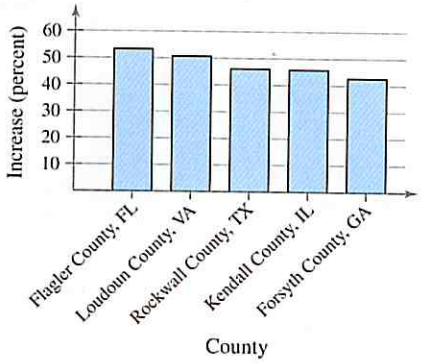
WHERE YOU'VE BEEN

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.

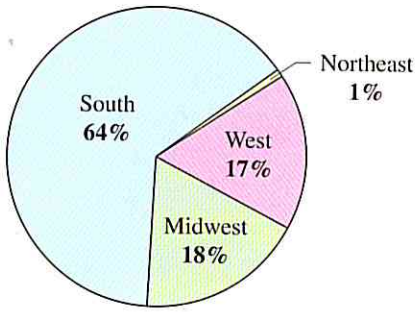
WHERE YOU'RE GOING

In Chapter 1, you will be introduced to the basic concepts and goals of statistics. For instance, statistics was used to construct the following graphs, which show the fastest growing U.S. counties with 10,000 or more people from 2000 to 2005 and the regions where the top 100 fastest growing counties from 2000 to 2005 are located. When conducting the census, the Bureau sends short forms to the entire population that ask about characteristics such as gender, age, race, and home ownership. A long form, which covers many additional topics, is sent to 17% of the population. This 17% forms 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. Counties



Location of the 100 Fastest Growing Counties



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 example, consider the following excerpts from recent newspapers and journals.

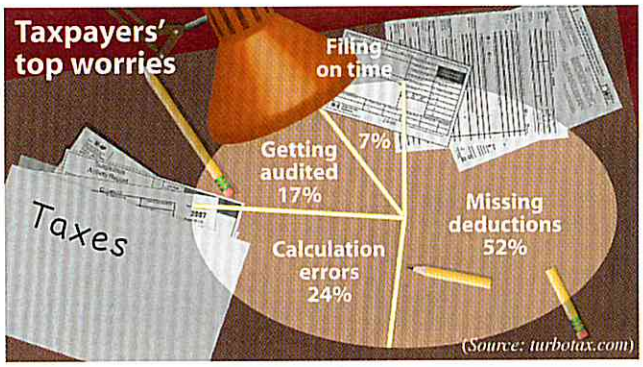
- "People who eat three daily servings of whole grains have been shown to reduce their risk of...stroke by 37%." *(Source: Whole Grains Council)*
- "Seventy percent of the 1500 U.S. spinal cord injuries to minors result from vehicle accidents, and 68 percent of those injured were not wearing a seatbelt." *(Source: UPI)*
- "U.S. coal production, which increased by 2.5 percent in 2006, is expected to fall by 3.1 percent in 2007." *(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.



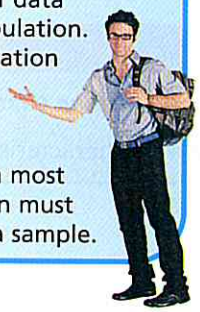
The use of statistics dates back to census taking in ancient Babylonia, Egypt, and later in the Roman Empire, when data were collected about matters concerning the state, such as births and deaths. In fact, the word *statistics* is derived from the Latin word *status,* meaning "state." So, what is statistics?

DEFINITION

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

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.



Data Sets

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

DEFINITION

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

A **sample** is a subset of a population.

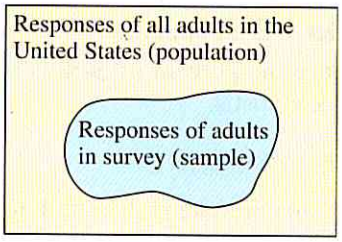
Sample data can be used to form conclusions about populations. Sample data must be collected using an appropriate method, such as *random selection*. If it is not collected using an appropriate method, the data are of no value.

EXAMPLE 1

Identifying Data Sets

In a recent survey, 1708 adults in the United States were asked if they think global warming is a problem that requires immediate government action. Nine hundred thirty-nine of the adults said yes. Identify the population and the sample. Describe the 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 1708 adults in the United States in the survey. The sample is a subset of the responses of all adults in the United States. The data set consists of 939 yes's and 769 no's.



Try It Yourself 1

The U.S. Department of Energy conducts weekly surveys of approximately 800 gasoline stations to determine the average price per gallon of regular gasoline. On February 12, 2007, the average price was $2.24 per gallon. Identify the population and the sample. *(Source: Energy Information Administration)*

a. Identify the *population*.
b. Identify the *sample*.
c. What does the data set consist of?

Answer: Page A32

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 newspaper.

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

Study Tip

The terms *parameter* and *statistic* are easy to remember if you use the mnemonic device of matching the first letters in *population parameter* and the first letters in *sample statistic*.

DEFINITION

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

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

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 a sample of MBAs reported that the average salary for an MBA is more than $82,000. *(Source: The Wall Street Journal)*
2. Starting salaries for the 667 MBA graduates from the University of Chicago Graduate School of Business increased 8.5% from the previous year.
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 $82,000 is based on a subset of the population, it is a sample statistic.
2. Because the percent increase of 8.5% is based on all 667 graduates' starting salaries, 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 2006, major league baseball teams spent a total of $2,326,706,685 on players' salaries. Does this numerical value describe a population parameter or a sample statistic? *(Source: USA Today)*

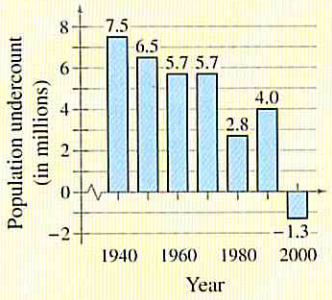
a. Decide whether the numerical value is from a *population* or a *sample*.
b. Specify whether the numerical value is a *parameter* or a *statistic*.

Answer: Page A32

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. This is an impossible task. It is important that the census be accurate, 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.

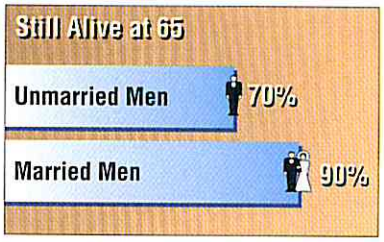
Inferential statistics is the branch of statistics that involves using a sample to draw conclusions about a population. A basic tool in the study of inferential statistics is probability.

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 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 A32

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.*

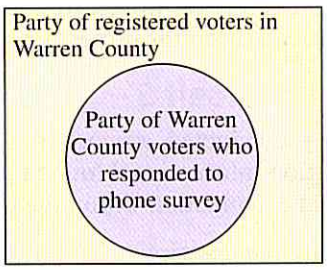
5. A statistic is a measure that describes a population characteristic.
6. A sample is a subset of a population.
7. It is impossible for the Census Bureau to obtain all the census data about the population of the United States.
8. Inferential statistics involves using a population to draw a conclusion about a corresponding sample.
9. A population is the collection of some outcomes, responses, measurements, or counts that are of interest.
10. The word *statistics* is derived from the Latin word *status*, meaning "state."

Classifying a Data Set *In Exercises 11–16, determine whether the data set is a population or a sample. Explain your reasoning.*

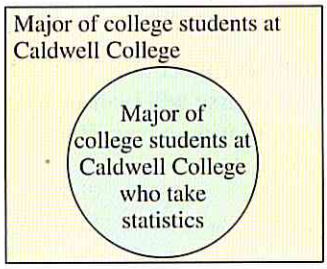
11. The age of each member of the House of Representatives
12. The height of every fourth person entering an amusement park
13. A survey of 500 spectators from a stadium with 42,000 spectators
14. The annual salary for each lawyer at a firm
15. The cholesterol levels of 20 patients in a hospital with 100 patients
16. The number of televisions in each U.S. household

Graphical Analysis *In Exercises 17–20, use the Venn diagram to identify the population and the sample.*

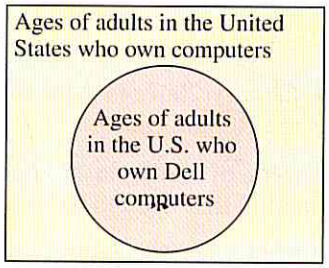
17.



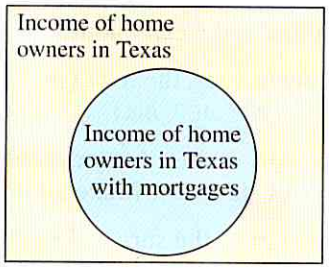
18.



19.



20.



Using and Interpreting Concepts

Identifying Populations and Samples *In Exercises 21–28, identify the population and the sample.*

21. A survey of 1000 adults in the United States found that 12% prefer to take their vacation during the winter months. *(Source: Rasmussen Reports)*

22. 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)*

23. A survey of 1906 households in the United States found that 13% have a high definition television.

24. A survey of 1000 computer users found that 17% plan on buying the Microsoft Windows Vista™ operating system. *(Source: Rasmussen Reports)*

25. A survey of 1045 registered voters found that 19% think the economy is the most important issue to consider when voting for Congress. *(Source: Princeton Survey Research Associates International)*

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

27. A survey of 546 women found that more than 56% are the primary investor in their household. *(Adapted from: Roper Starch Worldwide for Intuit)*

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

Distinguishing Between a Parameter and a Statistic *In Exercises 29–36, determine whether the numerical value is a parameter or a statistic. Explain your reasoning.*

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

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

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

32. As of January 2007, 44% of the governors of the 50 states in the United States are Republicans.

33. In a survey of a sample of computer users, 8% said their computer had a malfunction that needed to be repaired by a service technician.

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

35. In a recent survey of 1503 adults in the United States, 53% said they use both a landline and a cell phone. *(Source: Pew Research Center)*

36. In a recent year, the average math scores for all graduates on the ACT was 21.1. *(Source: ACT, Inc.)*

37. Which part of the survey described in Exercise 27 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

38. Which part of the survey described in Exercise 28 represents the descriptive branch of statistics? Make an inference based on the results of the survey.

Extending Concepts

39. 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?

40. Sleep Deprivation In a recent study, volunteers who had 8 hours of sleep were three times more likely to answer 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.

41. Living in Florida A study shows that senior citizens who live in Florida have a better memory 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?

42. 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?

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

(a) A study on the effectiveness of a new drug

(b) An analysis of a manufacturing process

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

Model	Base Price
Fusion 14 S	$17,795
F-150 XL	$18,710
Five Hundred SEL	$23,785
Escape XLT Sport	$24,575
2007 Explorer Sport Trac Limited	$26,775
Freestar SEL	$27,500
Crown Victoria LX	$28,830
Expedition XLT	$35,480

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 base prices of vehicle models. The names are nonnumerical entries, so these are qualitative data. The base 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
Cleveland, OH	452,208
Detroit, MI	886,671
Houston, TX	2,016,582
Las Vegas, NV	545,147
Portland, OR	533,427
Topeka, KS	121,946

a. Identify the contents of each data set.
b. Decide whether each data set consists of numerical or nonnumerical entries.
c. Specify the qualitative data and the quantitative data.

Answer: Page A32

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.*

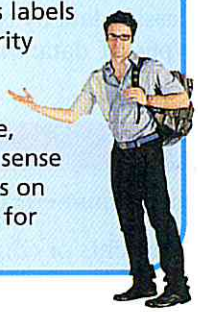
DEFINITION

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.

Insight

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.



PICTURING the WORLD

At the beginning of 2007, Forbes Magazine chose the 25 best U.S. cities to get a job. To gather the rankings, Forbes used five data points: unemployment rate, job growth, income growth, median household income, and cost of living. Growth data was then measured in the largest 100 metropolitan areas from 2003 to 2006. (Source: Forbes)

Forbes Top Five U.S. Cities for Jobs
1. Raleigh-Cary, NC
2. Phoenix-Mesa-Scottsdale, AZ
3. Jacksonville, FL
4. Orlando-Kissimmee, FL
5. Washington-Arlington-Alexandria, DC-VA-MD-WV

In this list, what is the level of measurement?

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: Nielsen Media Research)*

Top Five TV Programs (from 2/12/07 to 2/18/07)
1. American Idol–Tuesday
2. American Idol–Wednesday
3. Grey's Anatomy
4. House
5. CSI

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

Solution The first data set lists the rank of five TV programs. The data consist of the ranks 1, 2, 3, 4, and 5. Because the rankings 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 A32

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

DEFINITION

Data at the **interval level of measurement** can be ordered, and you can calculate meaningful differences between data entries. 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

2006 American League Home Run Totals (by Team)

Baltimore	164
Boston	192
Chicago	236
Cleveland	196
Detroit	203
Kansas City	124
Los Angeles	159
Minnesota	143
New York	210
Oakland	175
Seattle	172
Tampa Bay	190
Texas	183
Toronto	199

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

$$2000 - 1923 = 77 \text{ years.}$$

But it does not make sense to say that one year is a multiple of another. So, these data are at the interval level. Using the home run totals, you can find differences *and* write ratios. From the data, you can see that Detroit hit 31 more home runs than Seattle hit and that Chicago hit about twice as many home runs as Kansas City 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 A32

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 Music Played by a Radio Station* Pop Modern rock Contemporary jazz Hip hop	*Put in a category.* For instance, a song played by the radio station could be put into one of the four 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 Temperature (in degrees Fahrenheit) for Sacramento, CA* Jan 46.3, Feb 51.2, Mar 54.5, Apr 58.9, May 65.5, Jun 71.5, Jul 75.4, Aug 74.8, Sep 71.7, Oct 64.4, Nov 53.3, Dec 45.8 *(Source: National Climatic Data Center)*	Put in a category, put in order, and *find differences between values.* For instance, 71.5 − 65.5 = 6°F. So, June is 6° warmer than May.
Ratio Level (Quantitative data)	*Average Monthly Precipitation (in inches) for Sacramento, CA* Jan 3.8, Feb 3.5, Mar 2.8, Apr 1.0, May 0.5, Jun 0.2, Jul 0.1, Aug 0.1, Sep 0.4, Oct 0.9, Nov 2.2, Dec 2.5 *(Source: National Climatic Data Center)*	Put in a category, put in order, find differences between values, and *find ratios of values.* For instance, $\frac{1.0}{0.5} = 2$. So, there is twice as much rain in April 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.

Classifying Data by Type *In Exercises 7–12, determine whether the data are qualitative or quantitative.*

7. The telephone numbers in a telephone directory
8. The daily high temperatures for the month of July
9. The lengths of songs on an MP3 player
10. The player numbers for a soccer team
11. Responses on an opinion poll
12. Measure of diastolic blood pressure

Using and Interpreting Concepts

Classifying Data by Level *In Exercises 13–18, determine whether the data are qualitative or quantitative and identify the data set's level of measurement. Explain your reasoning.*

13. **Football** The top five teams in the final college football poll released in January 2007 are listed. *(Source: Associated Press)*

    1. Florida 2. Ohio State 3. LSU 4. USC 5. Boise State

14. **Politics** The three political parties in the 110th Congress are listed below.

    Republican Democrat Independent

15. **Top Salesperson** The region representing the top salesperson in a corporation for the past six years is given.

    Southeast Northwest Northeast
    Southeast Southwest Southwest

16. **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

17. Best Seller List The top five hardcover fiction books on *The New York Times* Best Seller List on February 21, 2007 are shown. *(Source: The New York Times)*

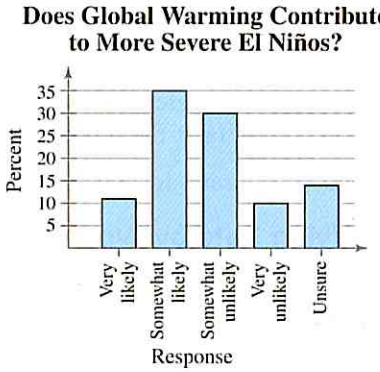
1. Step on a Crack
2. Plum Lovin'
3. Natural Born Charmer
4. High Profile
5. Hannibal Rising

18. Ticket Prices The average ticket prices for 10 rock concerts in 2005 are listed. *(Source: The New York Times)*

$134 $104 $55 $63 $76 $38 $35 $81 $47 $97

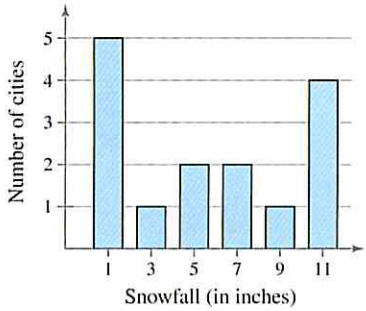
Graphical Analysis *In Exercises 19–22, identify the level of measurement of the data listed on the horizontal axis in the graph.*

19.



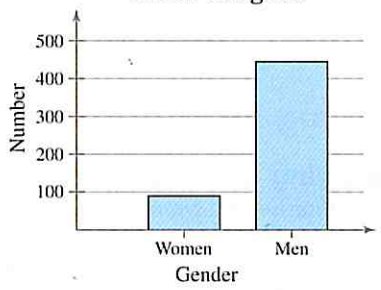
(Source: Yankelovich for the National Representatives Science Foundation, American Meteorological Society)

20.



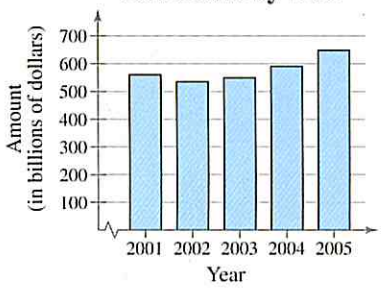
(Source: National Climatic Data Center)

21.



(Source: U.S. House of Representatives, Office of the Clerk)

22.



(Source: U.S. Census Bureau)

23. The following 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)

24. The following 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

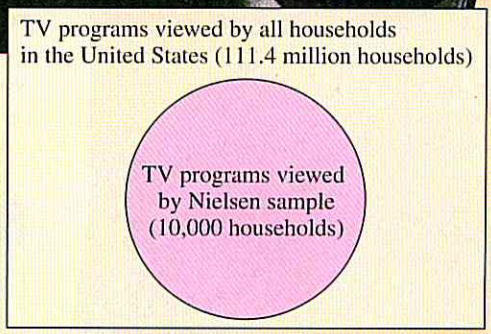
25. Writing What is an inherent zero? Describe three examples of data sets that have an inherent zero and three that do not.

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

CASE STUDY

Rating Television Shows in the United States

Nielsen Media Research has been rating television programs for more than 50 years. Nielsen uses several sampling procedures, but its main one is to track the viewing patterns of 10,000 households. These contain more than 30,000 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 10,000 households are used to draw inferences about the population of all households in the United States.



Top-Ranked Programs in Prime Time for the Week of 2/12/07–2/18/07

Rank	Rank Last Week	Program Name	Network	Day, Time	Rating	Share	Audience
1	1	American Idol–Tuesday	FOX	Tues., 8:00 P.M.	17.4	25	19,354,000
2	2	American Idol–Wednesday	FOX	Wed., 9:00 P.M.	16.2	24	18,045,000
3	3	Grey's Anatomy	ABC	Thu., 9:00 P.M.	16.0	23	17,809,000
4	4	House	FOX	Tues., 9:00 P.M.	14.8	22	16,469,000
5	5	CSI	CBS	Thu., 9:00 P.M.	13.8	20	15,323,000
6	7	CSI: Miami	CBS	Mon., 10:00 P.M.	12.7	21	14,093,000
7	8	Desperate Housewives	ABC	Sun., 9:00 P.M.	11.7	18	13,060,000
8	10	Deal or No Deal–Monday	NBC	Mon., 8:00 P.M.	10.0	15	11,167,000
8	8	Two and a Half Men	CBS	Mon., 9:00 P.M.	10.0	14	11,099,000
10	17	Shark	CBS	Thu., 10:00 P.M.	9.8	16	10,909,000

Exercises

1. **Rating Points** Each rating point represents 1,114,000 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? What is the unit of measure for the difference of two entries in the data set?
6. **Ratio Level of Measurement** Which three columns contain data at the ratio level?
7. **Share** The column listed as "Share" gives the percentage of televisions in use at a given time. Does the Nielsen rating rank shows by rating or by share? Explain your reasoning.
8. **Inferences** What decisions (inferences) can be made on the basis of the Nielsen ratings?

1.3 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. 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

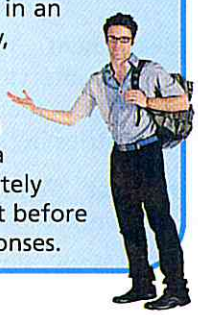
There are several ways you can collect data. Often, the focus of the study dictates the best way to collect data. The following is a brief summary of four methods of data collection.

- ***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. 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

The difference between an observational study and an experiment is that, in an observational study, a researcher does not influence the responses, whereas in an experiment, a researcher deliberately applies a treatment before observing the responses.



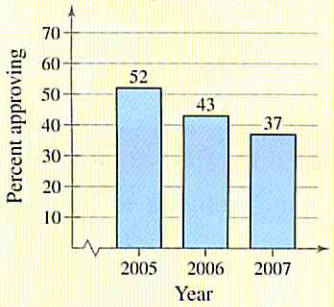
- ***Use a simulation*** A **simulation** is the use of a mathematical or physical model to reproduce the conditions of a situation or process. Collecting data often involves the use of computers. Simulations allow you to study situations that are impractical or even dangerous to create in real life, 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. For example, 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 example, the approval ratings for President George W. Bush from 2005 to 2007 are shown in the following graph. (The rating is from the first poll conducted in January of each year.)

President's Approval Ratings, 2005–2007



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 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 A32

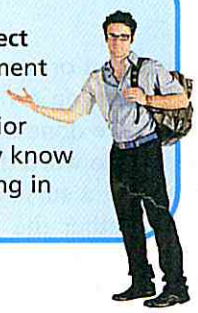
Experimental Design

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 in important. One such factor is a *confounding variable*.

Insight

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



DEFINITION

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

For example, 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, he or she has been given no medicated treatment at all. To help control or minimize the placebo effect, a technique called **blinding** can be used.

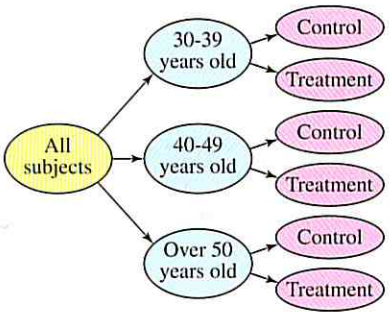
DEFINITION

Blinding is a technique where the subject does not know whether he or she is receiving a treatment or a placebo. In a **double-blind experiment,** neither the subject nor the experimenter knows if the subject is 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 example, 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. Then within each age group, randomly assign subjects to either the treatment group or control group as shown.

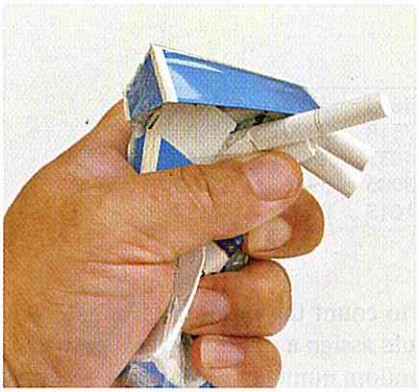
Another type of experimental design is a **matched-pairs design,** where subjects are paired up according to a similarity. One subject in the pair is randomly selected to receive one treatment while the other subject receives a different treatment. For instance, two subjects may be paired up because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

Another important part of experimental design is the sample size. To improve the validity of experimental results, *replication* is required.

DEFINITION

Replication is the repetition of an experiment using a large group of subjects.

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 both groups were 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, the female group has a significant number of subjects who 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 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 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 A32

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.

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 these 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 from 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 36).

To explore this topic further, see Activity 1.3 on page 28.

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

For instance, 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 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

Using a Simple Random Sample

There are 731 students currently enrolled in statistics 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 each student in the course. On 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 example, 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, follow the instructions in the margin.

Try It Yourself 3

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

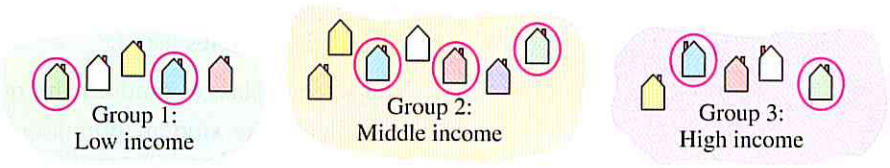
a. On the table, randomly choose a *starting place*.
b. *Read the digits* in groups of two.
c. Write the five random numbers.

Answer: Page A32

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.

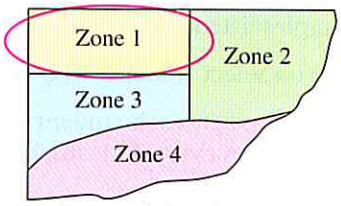
- ***Stratified Sample*** When it is important for the sample to have members from each segment of the population, you should use a stratified sample. Depending on the focus of the study, members of the population are divided into two or more subsets, called *strata*, that share a similar characteristic 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 example, 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 example, if one of the zip code clusters has a greater proportion of high-income people, the data might not be representative of the population.

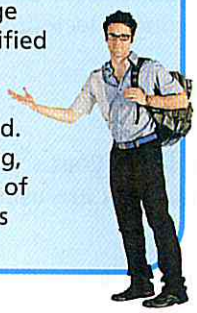
Zip Code Zones in West Ridge County



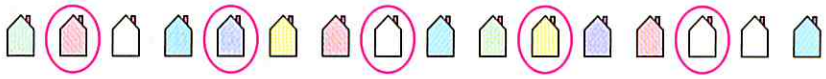
Cluster Sampling

Insight

For stratified sampling, each of the strata contains members with a certain characteristic (for example, a particular age group). In contrast, clusters consist of geographic groupings, and each cluster should consist of members with all of the characteristics (for example, 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.



- ***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 example, 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 opinion of students at your school 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 divide the student population with respect to majors and randomly select and question some students in each major.
3. You assign each student a number and generate random numbers. You then question each student whose number is randomly selected.

Solution

1. Because each class is a naturally occurring subgroup (a cluster) and you question each student in the class, this is a cluster sample.
2. Because students are divided into strata (majors) and a sample is selected from each major, this is a stratified sample.
3. 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.

Try It Yourself 4

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

1. You select students who are in your statistics 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.*
b. Identify the corresponding *sampling technique.*

Answer: Page A32

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. Describe two methods you can use to generate random numbers.
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 11–14, decide which method of data collection you would use to collect data for the study. Explain.*

11. A study of the effect on the human digestive system of potato chips made with a fat substitute
12. A study of the effect of a product's warning label to determine whether consumers still buy the product
13. A study of how fast a virus would spread in a metropolitan area
14. A study of the ages of the 535 members of the U.S. Congress

Using and Interpreting Concepts

15. **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 in this experiment.

    (b) How many treatments are used in this experiment?

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

    (d) How could this experiment be designed to be double-blind?

16. 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 sneaker 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 an MRI was performed on their knees. *(Source: Washington Post)*

(a) Identify the experimental units in this experiment.

(b) How many treatments are used in this experiment?

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

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

(e) 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 17–26, identify the sampling technique used and discuss potential sources of bias (if any). Explain.*

17. Using random digit dialing, researchers called 1599 people and asked what obstacles (such as childcare) kept them from exercising. *(Source: Yankelovich Partners, Inc. for Shape Up America!)*

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

19. Questioning students as they left a university library, a researcher asked 358 students about their drinking habits.

20. 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.

21. Chosen at random, 1210 hospital outpatients were contacted and asked their opinion of the care they received.

22. For quality assurance, every twentieth engine part is selected from an assembly line and tested for durability.

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

24. Questioning teachers as they left a faculty lounge, a researcher asked 32 teachers about their teaching styles and grading methods.

25. A list of managers is compiled and ordered. After a starting number is randomly chosen, every ninth name is selected until 1000 managers are selected. The managers are questioned about the use of digital media.

26. From calls made with randomly generated telephone numbers, 1012 respondents were asked if they rented or owned their residence.

27. 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 9 subjects for the treatment group. The other 9 subjects will go into the control group. List the subjects in each group. Tell which method you used to generate the random numbers.

28. 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 27 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 29 and 30, 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.*

29. The average salary of the 50 employees of a company

30. The most popular car color among 25,000 students at a university

Recognizing a Biased Question *In Exercises 31–34, determine whether the survey question is biased. If the question is biased, suggest a better wording.*

31. Why is drinking fruit juice good for you?

32. Why are drivers who change lanes several times dangerous?

33. How many hours of sleep do you get on an average night?

34. Do you think the media have a negative effect on teen girls' dieting habits?

35. Writing Television program ratings by Nielsen Media Research are described on page 17. Discuss the strata used in the sample.

36. Writing Television program ratings by Nielsen Media Research are described on page 17. Why is it important to have a stratified sample for these ratings?

Extending Concepts

37. 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 only for a fixed response. An open question and a closed question with its possible choices are given below. List an advantage and a disadvantage of an open question. Then list an advantage and a disadvantage of a closed 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

38. 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?

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

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

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

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

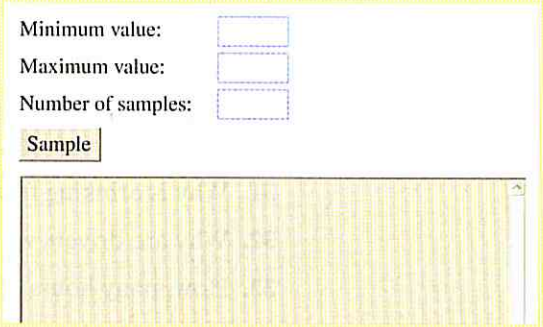
43. Using Sampling Techniques You have been asked by your school to survey 150 students who use the new fitness center. Describe your procedure for obtaining a sample of each type: random, stratified, cluster, systematic, and convenience.

Random Numbers

ACTIVITY 1.3

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.



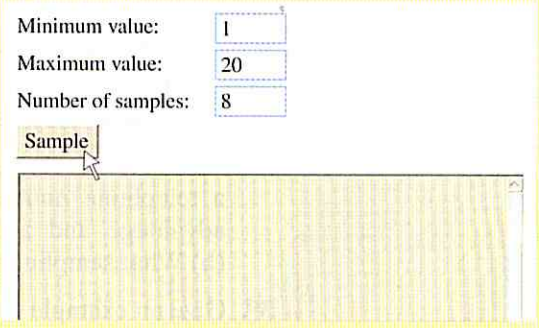
Explore

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?



2. Use the applet to repeat Example 3 on page 22. What values did you use for the minimum, maximum, and number of samples? Which method do you prefer? Explain.

Uses & 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 enough 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 experiment would not want to deny the new drug for 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 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*	*11–14*
How to classify data with respect to the four levels of measurement: nominal, ordinal, interval, and ratio	*2, 3*	*15–18*
Section 1.3		
How data are collected: by doing an observational study, performing an experiment, using a simulation, or using a survey	*1*	*19–22*
How to design an experiment	*2*	*23, 24*
How to create a sample using random sampling, simple random sampling, stratified sampling, cluster sampling, and systematic sampling	*3, 4*	*25–30*
How to identify a biased sample	*3, 4*	*31–34*

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 92% are worried about dependence on foreign oil. *(Source: Yale University)*
2. Thirty-eight nurses working in the San Francisco area were surveyed concerning their opinions of managed health care.
3. A study of 146 credit cards found that the average late fee is $27.46. *(Source: Consumer Action)*
4. A survey of 1205 physicians found that about 60% 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 2006 team payroll of the New York Mets was $101,084,963. *(Source: U.S.A. 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. Nineteen percent of a sample of Indiana ninth graders surveyed said they smoked cigarettes daily. *(Source: Indiana University)*
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 11–14, 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

In Exercises 15–18, identify the data set's level of measurement. Explain your reasoning.

15. 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

16. The EPA size classes for a sample of automobiles are listed.

    subcompact compact midsize large compact large

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

Administration Sales Production Billing

18. The heights (in inches) of the 2006 Los Angeles Sparks are listed. *(Source: Women's National Basketball Association)*

69 74 63 77 71 74 75 70 74 75 75 75 71

Section 1.3

In Exercises 19–22, decide which method of data collection you would use to collect data for the study. Explain your reasoning.

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

20. A study of the effect of kangaroos on the Florida Everglades ecosystem

21. A study of the effects of a fertilizer on a soybean crop

22. A study of college students' opinions on environmental pollution

In Exercises 23 and 24, 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.

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

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

In Exercises 25–30, identify which sampling technique was used in the study. Explain your reasoning.

25. Calling randomly generated telephone numbers, a study asked 1001 U.S. adults which medical conditions could be prevented by their diet. *(Adapted from Wirthlin Worldwide)*

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

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

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

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

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

In Exercises 31–34, identify a bias or error that might occur in the indicated survey or study.

31. study in Exercise 25

32. experiment in Exercise 26

33. study in Exercise 27

34. sampling in Exercise 28

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 372 patients with anxiety disorders was conducted to find a link between coffee drinking and anxiety disorders.

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

(a) In a survey of 798 Internet users, 19% said they have a wireless network in their home. *(Source: Pew Internet and American Life Project)*

(b) In a vote, 84% of the employees at a company voted for new vending machines in the building.

(c) A survey of about 1000 Americans shows that only 40% have an emergency savings account. *(Source: Consumer Federation of America)*

3. Determine whether the data are qualitative or quantitative.

(a) A list of post office box numbers

(b) The final exam scores in a chemistry class

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 number of candles sold by a candle manufacturer each quarter for the current fiscal year

(c) 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 to participate 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 which sampling technique was 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 in fields 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.

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 chose 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) 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.

How did you acquire your ticket?

Response	Percent
Travel agent	35.1%
Directly from airline	20.9%
Online, using the airline's Web site	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)