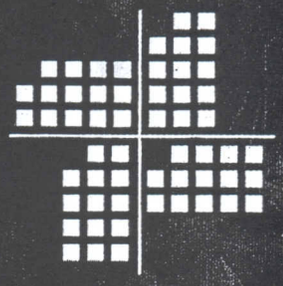


1 • The Role of Statistics

Statistical methods for summary and analysis of data provide investigators with powerful tools for making sense out of data. Statistical techniques are being used with increasing frequency in business, medicine, agriculture, social sciences, natural sciences, and applied sciences, such as engineering. The pervasiveness of statistical analyses in such diverse fields has led to increased recognition that statistical literacy — a familiarity with the goals and methods of statistics — should be a basic component of a well-rounded educational program. In this chapter, we consider the nature and role of variability in statistical settings, introduce some basic terminology, and look at some simple graphical displays for summarizing data.

▪ 1.1 Three Reasons to Study Statistics

Because of the widespread use of statistical methods to organize, summarize, and draw conclusions from data, a familiarity with statistical techniques and statistical literacy in general is vital in today's society. It's a good idea for everyone to have a basic understanding of statistics, and many college majors require at least one course in statistics. There are three important reasons for this: (1) to be informed, (2) to understand issues and be able to make decisions, and (3) to be able to evaluate decisions that affect your life. Let's explore each reason in detail.

▪ The First Reason: Being Informed

In today's society, we are bombarded with numerical information in news, in advertisements, and even in conversation. How do we decide whether claims based on numerical information are reasonable? Consider the examples that appeared in one week's news (week of October 23, 2002).

- An analysis of data from a University of Utah study led to the conclusion that drivers engaged in cell phone conversations missed twice as many simulated

traffic signals as drivers who were not talking and that cell phone users took longer to react to the signals they did detect. The researchers also found that these "driving deficits" were the same whether drivers used hands-free devices or held the phone to the ear. (*Los Angeles Times*, October 23, 2002)

- The Food and Drug Administration warned consumers about serious risks associated with the use of decorative contact lenses. These novelty lenses are imprinted with designs ranging from holiday decorations to sports logos. By analyzing data on the occurrence of corneal abrasions and other eye injuries, the FDA determined that use of these contact lenses, which generally have not been properly prescribed and fitted, can jeopardize vision. (*USA Today*, October 23, 2002)
- Results from a long-term study of cancer rates in people vaccinated for polio were reported. Between 1955 and 1963, about one-third of those vaccinated for polio received a vaccine that was contaminated with a virus called SV40. Recent experiments have shown that SV40 causes cancer and that SV40 has been detected in some types of rare cancer tumors. Researchers inferred that there has been no increase in cancer among people who received the contaminated vaccine, but they called for more research because the tumors associated with the SV40 virus are so rare that more information is needed to reach a definitive conclusion. (*USA Today*, October 23, 2002)
- Data on alcohol, tobacco, and marijuana use by junior and senior high school students in San Luis Obispo, California, were summarized. Graphs were used to show how the proportions of students who reported using each of the three substances increase from 7th to 11th grade and to confirm that these proportions did not change much between 1999 and 2001. (*San Luis Obispo Tribune*, October 25, 2002)
- In a summary of the "Human Footprint" report, the Wildlife Conservation Society and Columbia University described an analysis of the Earth's land surface and concluded that 83% of the land surface is used by humans for housing, farming, mining, or fishing. Among the few remaining wild areas are the northern forests of Alaska, Canada, and Russia, the high plateaus of Tibet, and parts of the Amazon River basin. (*USA Today*, October 23, 2002)
- A study that links hospital patient care and subsequent well-being to nurse staffing level was summarized in the *Journal of the American Medical Association*. Based on a survey of more than 10,000 nurses at 168 Pennsylvania hospitals and an analysis of the medical records of 232,342 patients who underwent routine surgery at these hospitals, the investigators concluded that surgery patients at hospitals with a severe shortage of nurses had a 31% greater risk of dying while in the hospital. (*USA Today* and *Los Angeles Times*, October 23, 2002)
- Water quality was compared for beaches in three Southern California counties. Based on data from water specimens taken at the beaches, the beaches were graded from A+ to F according to the risk of getting sick from swimming. It was reported that 70% of the 106 beaches in Orange County, 72% of the 82 beaches in Los Angeles County, and 90% of the 50 beaches in Ventura County received A grades. (*Los Angeles Times*, October 25, 2002)
- It was reported that, contrary to popular belief, both men and women tend to get more jealous over sexual rather than emotional infidelity. Previous studies

had found that men tended to say it would be more upsetting to find out that their partners had been unfaithful, whereas women were more likely to say it would be more upsetting if their partner formed a strong emotional bond with someone else. However, in a new study involving 111 subjects, both men and women rated sexual infidelity as most upsetting, suggesting a change in attitudes over time. (*San Luis Obispo Tribune*, October 29, 2002)

- A study concluded that giving aspirin to heart patients soon after bypass surgery dramatically lowers the risk of death and complications. This conclusion was based on an experiment in which about 60% of 5065 patients who underwent bypass surgery received aspirin within 48 hours of surgery. The researchers found that those receiving aspirin were less likely to die in the hospital and less likely to suffer a heart attack, stroke, or kidney failure while in the hospital. (*San Luis Obispo Tribune*, October 24, 2002)

To be an informed consumer of such reports, you must be able to (1) extract information from charts and graphs, (2) follow numerical arguments, and (3) know the basics of how data should be gathered, summarized, and analyzed to draw statistical conclusions.

▪ The Second Reason: Understanding and Making Decisions

No matter what profession you choose, you will almost certainly need to understand statistical information and base decisions on it. Here are some examples:

- Almost all industries, as well as government and nonprofit organizations, use market research tools, such as consumer surveys, that are designed to provide information about who uses their products or services.
- Modern science and its applied fields, from astrophysics to zoology, rely on statistical methods for analyzing data and deciding whether various conjectures are supported by observed data. This is also true for the social sciences, such as economics and psychology. Even the liberal arts fields, such as literature and history, are beginning to use statistics as a research tool.
- In law or government, you may be called on to understand and debate statistical techniques used in another field. Class-action lawsuits can depend on a statistical analysis of whether one kind of injury or illness is more common in a particular group than in the general population. Proof of guilt in a criminal case may rest on statistical interpretation of the likelihood that DNA samples match.

Throughout your professional life, you will have to make informed decisions. To make these decisions, you must be able to do the following:

1. Decide whether existing information is adequate or whether additional information is required.
2. If necessary, collect more information in a reasonable and thoughtful way.
3. Summarize the available data in a useful and informative manner.
4. Analyze the available data.
5. Draw conclusions, make decisions, and assess the risk of an incorrect decision.

Cognitive psychologists have shown that people informally use these steps to make everyday decisions. Should you go out for a sport that involves the risk of

injury? Will your college club do better by trying to raise funds with a benefit concert or with a direct appeal for donations? If you choose a particular major, what are your chances of finding a job when you graduate? How should you select a graduate program based on guidebook ratings that include information on percentage of applicants accepted, time to obtain a degree, and so on? The study of statistics formalizes the process of making decisions based on data and provides the tools for accomplishing the steps listed.

▪ The Third Reason: Evaluating Decisions That Affect Your Life

It is likely that you will need to make decisions based on data. Other people also use statistical methods to make decisions that affect you as an individual. An understanding of statistical techniques will allow you to question and evaluate decisions that affect your well-being. Some examples are:

- Insurance companies use statistical techniques to set auto insurance rates, although some states restrict the use of these techniques. Data suggest that young drivers have more accidents than older ones. Should laws or regulations limit how much more young drivers pay for insurance? What about the common practice of charging higher rates for people who live in urban areas?
- University financial aid offices survey students on the cost of going to school and collect data on family income, savings, and expenses. The resulting data are used to set criteria for deciding who receives financial aid. Are the estimates they use accurate?
- Medical researchers use statistical methods to make recommendations regarding the choice between surgical and nonsurgical treatment of such diseases as coronary heart disease and cancer. How do they weigh the risks and benefits to reach such a recommendation?
- Many companies now require drug screening as a condition of employment, but with these screening tests there is a risk of a false-positive reading (incorrectly indicating drug use) or a false-negative reading (failure to detect drug use). What are the consequences of a false result? Given the consequences, is the risk of a false result acceptable?

An understanding of elementary statistical methods can help you to evaluate whether important decisions such as the ones just mentioned are being made in a reasonable way.

We encounter data and conclusions based on data every day. **Statistics** is the scientific discipline that provides methods to help us make sense of data. Some people are suspicious of conclusions based on statistical analyses. Extreme skeptics, usually speaking out of ignorance, characterize the discipline as a subcategory of lying —something used for deception rather than for positive ends. However, we believe that statistical methods, used intelligently, offer a set of powerful tools for gaining insight into the world around us. We hope that this textbook will help you to understand the logic behind statistical reasoning, prepare you to apply statistical methods appropriately, and enable you to recognize when others are not doing so.

▪ 1.2 The Nature and Role of Variability

Statistics is the science of collecting, analyzing, and drawing conclusions from data. If we lived in a world where all measurements were identical for every individual, all three of these tasks would be simple. Imagine a population consisting of all students at a particular university. Suppose that *every* student took the same number of units, spent exactly the same amount of money on textbooks this semester, and favored increasing student fees to support expanding library services. For this population, there is *no* variability in the values of number of units, amount spent on books, or student opinion on the fee increase. A researcher studying a sample from this population to draw conclusions about these three variables would have a particularly easy task. It would not matter how many students the researcher included in the sample or how the sampled students were selected. In fact, the researcher could collect information on number of units, amount spent on books, and opinion on the fee increase by just stopping the next student who happened to walk by the library. Because there is no variability in the population, this one individual would provide complete and accurate information about the population, and the researcher could draw conclusions based on the sample with no risk of error.

The situation just described is obviously unrealistic. Populations with no variability are exceedingly rare, and they are of little statistical interest because they present no challenge! In fact, variability is almost universal. It is variability that makes life (and the life of a statistician, in particular) interesting. We need to understand variability to be able to collect, analyze, and draw conclusions from data in a sensible way. One of the primary uses of descriptive statistical methods is to increase our understanding of the nature of variability in a population.

Examples 1.1 and 1.2 illustrate how an understanding of variability is necessary to draw conclusions based on data.

▪ Example 1.1 If the Shoe Fits

The graphs in Figure 1.1 are examples of a type of graph called a histogram. (We will see how to construct histograms in Chapter 3.) Figure 1.1(a) shows the distribution of the heights of female basketball players who played at a particular university between 1990 and 1998. The height of each bar in the graph indicates how many players' heights were in the corresponding interval. For example, 40 basketball players had heights between 72 in. and 74 in., whereas only 2 players had heights between 66 in. and 68 in. Figure 1.1(b) shows the distribution of heights for members of the women's gymnastics team over the same period. Both histograms are based on the heights of 100 women.

The first histogram shows that the heights of female basketball players varied, with most heights falling between 68 in. and 76 in. In the second histogram we see that the heights of female gymnasts also varied, with most heights in the range of 60 in. to 72 in. It is also clear that there is more variation in the heights of the gymnasts than in the heights of the basketball players, because the gymnast histogram spreads out more about its center than does the basketball histogram.

Now suppose that a tall woman (5 ft 11 in.) tells you she is looking for her sister who is practicing with her team at the gym. Would you direct her to where the basketball team is practicing or to where the gymnastics team is practicing? What

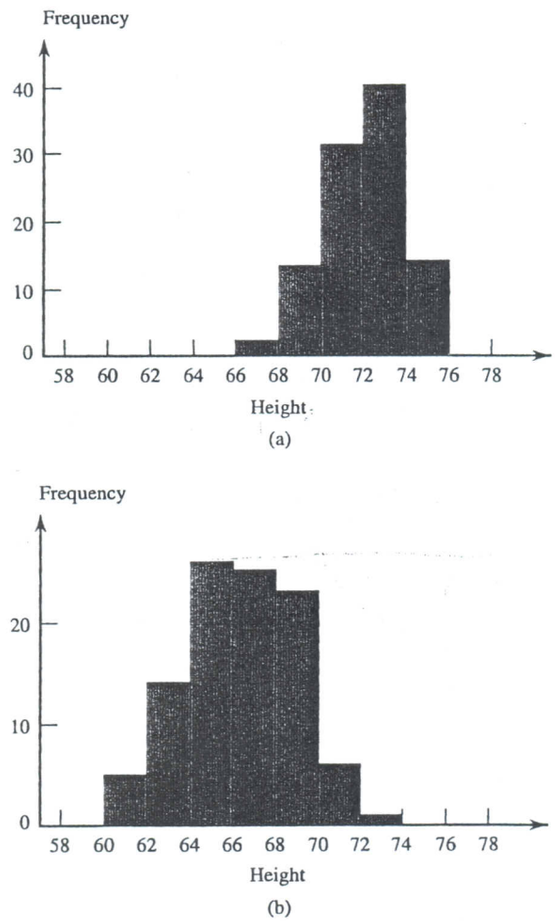


Figure 1.1 Histograms of heights (in inches) of female athletes: (a) basketball players; (b) gymnasts.

reasoning would you use to decide? What if you found a pair of size 6 tennis shoes left in the locker room? Would you first try to return them by checking with members of the basketball team or the gymnastics team?

You probably answered that you would send the woman looking for her sister to the basketball practice and that you would try to return the shoes to a gymnastics team member. To reach these conclusions, you informally used statistical reasoning that combined your own knowledge of the relationship between heights of siblings and between shoe size and height with the information about the distributions of heights presented in Figure 1.1. You might have reasoned that heights of siblings tend to be similar and that a height as great as 5 ft 11 in., although not impossible, would be unusual for a gymnast. On the other hand, a height as tall as 5 ft 11 in. would be a common occurrence for a basketball player. Similarly, you might have reasoned that tall people tend to have bigger feet and that short people tend to have smaller feet. The shoes found were a small size, so it is more likely that they belong to a gymnast than to a basketball player, because small heights and small feet are usual for gymnasts and unusual for basketball players.

▪ Example 1.2 Monitoring Water Quality

As part of its regular water quality monitoring efforts, an environmental control board selects five water specimens from a particular well each day. The concentration of contaminants in parts per million (ppm) is measured for each of the five specimens, and then the average of the five measurements is calculated. The histogram in Figure 1.2 summarizes the average contamination values for 200 days.

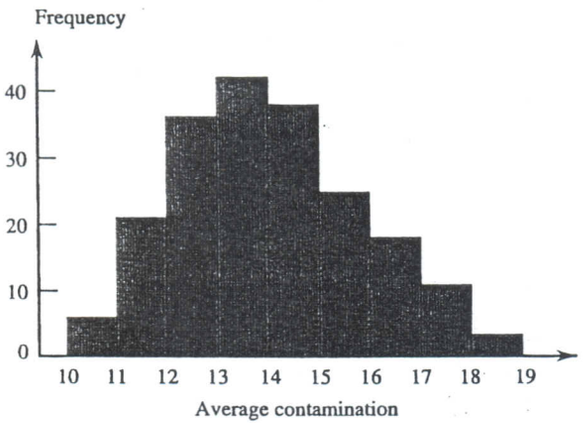


Figure 1.2 Contaminant concentration (in parts per million) in well water.

Now suppose that a chemical spill has occurred at a manufacturing plant about 1 mile from the well. It is not known whether a spill of this nature would contaminate groundwater in the area of the spill and, if so, whether a spill this distance from the well would affect the quality of well water.

One month after the spill, five water specimens are collected from the well, and the average contamination is 16 ppm. Considering the variation before the spill, would you take this as convincing evidence that the well water was affected by the spill? What if the calculated average was 18 ppm? 22 ppm? How is your reasoning related to the graph in Figure 1.2?

Before the spill, the average contaminant concentration varied from day to day. An average of 16 ppm would not have been an unusual value, and so seeing an average of 16 ppm after the spill isn't necessarily an indication that contamination has increased. On the other hand, an average as large as 18 ppm is less common, and an average as large as 22 ppm is not at all typical of the prespill values. In this case, we would probably conclude that the well contamination level has increased.

In these two examples, reaching a conclusion required an understanding of variability. Understanding variability allows us to distinguish between usual and unusual values. The ability to recognize unusual values in the presence of variability is the essence of most statistical procedures and is also what enables us to quantify the chance of being incorrect when a conclusion is based on sample data. These concepts will be developed further in subsequent chapters.

▪ 1.3 Statistics and Data Analysis

Statistical methods, used appropriately, allow us to draw reliable conclusions based on data. Data and conclusions based on data appear regularly in a variety of settings: newspapers, advertisements, magazines, and professional journals. In business, industry, and government, informed decisions are often data driven.

Statistics is the science of collecting, analyzing, and drawing conclusions from data.

Once data have been collected or once an appropriate data source has been identified, the next step in the data analysis process usually involves organizing and summarizing the information. Tables, graphs, and numerical summaries allow increased understanding and provide an effective way to present data. Methods for organizing and summarizing data make up the branch of statistics called **descriptive statistics.**

After the data have been summarized, we often wish to draw conclusions or make decisions based on the data. This usually involves generalizing from a small group of individuals or objects that we have studied to a much larger group.

For example, the admissions director at a large university might be interested in learning why some applicants who were accepted for the fall 2004 term failed to enroll at the university. The population of interest to the director consists of all accepted applicants who did not enroll in the fall 2004 term. Because this population is large and because it may be difficult to contact all the individuals, the director might be able to collect data from only 300 selected students. These 300 students constitute a sample.

▪ **Definition**

The entire collection of individuals or objects about which information is desired is called the **population** of interest. A **sample** is a subset of the population, selected for study in some prescribed manner.

The second major branch of statistics, **inferential statistics**, involves generalizing from a sample to the population from which it was selected. When we generalize in this way, we run the risk of an incorrect conclusion, because a conclusion about the population is based on incomplete information. An important aspect in the development of inferential techniques involves quantifying the chance of an incorrect conclusion.

Considering some examples will help you develop a preliminary appreciation for the scope and power of statistical methods. In Examples 1.3–1.5, we describe three problems that can be investigated using techniques presented in this textbook.

▪ Example 1.3 Student Opinion Survey

A university has recently implemented a new registration system. Students interact online with the computer to select classes for the term. To assess student opinion regarding the effectiveness of the system, a university research team wants to conduct a survey. Each student in a sample of 400 will be asked a variety of questions (such as the number of units received and the number of times the system was accessed before the registration process was completed). The survey will yield a rather large and unwieldy data set. To make sense out of the raw data and to describe student responses, the researchers must summarize the data. This would also make the results more accessible to others. Descriptive techniques can be used to accomplish this task. In addition, inferential methods can be employed to draw various conclusions about the experiences of *all* students who used the registration system.

▪ Example 1.4 Depression and Cholesterol Levels

A study linking depression to low cholesterol levels was described in an Associated Press article (*San Luis Obispo Telegram Tribune*, June 23, 1995). Researchers at a hospital in Italy compared the average cholesterol level for a sample of 331 patients who had been admitted to the hospital after a suicide attempt and who had been diagnosed with clinical depression to the average cholesterol level of 331 patients admitted to the hospital for other reasons. Statistical techniques were used to analyze the data and to show that the average cholesterol level was lower for the depressed group. The article correctly noted that because of the way in which the data were collected, it was not possible to determine from the statistical analysis alone whether a causal relationship between cholesterol level and psychological state exists — that is, whether low cholesterol levels affect psychological state or vice versa.

▪ Example 1.5 Predicting the Spread of a Forest Fire

A final example comes from the discipline of forestry. When a fire occurs in a forested area, decisions must be made about the best way to combat the fire. One possibility is to try to contain the fire by building a fire line. If building a fire line requires 4 hr, deciding where the line should be built involves making a prediction of how far the fire will spread during this period. Many factors must be taken into account, including wind speed, temperature, humidity, and time elapsed since the last rainfall. Statistical techniques make it possible to develop a model for the prediction of fire spread, using information available from past fires.

▪ Exercises 1.1–1.7

1.1 Give a brief definition of the terms *descriptive statistics* and *inferential statistics*.

1.2 Give a brief definition of the terms *population* and *sample*.

1.3 The student senate at a university with 15,000 students is interested in the proportion of students who favor a change in the grading system to allow for plus and minus grades (e.g., B+, B, B−, rather than just B). Two hundred students are interviewed to determine their attitude toward this proposed change. What is the population of interest? What group of students constitutes the sample in this problem?

1.4 The supervisors of a rural county are interested in the proportion of property owners who support the construction of a sewer system. Because it is too costly to contact all 7000 property owners, a survey of 500 owners (selected at random) is undertaken. Describe the population and sample for this problem.

1.5 Representatives of the insurance industry wished to investigate the monetary loss resulting from earthquake damage to single-family dwellings in Northridge, California, in January 1994. From the set of all single-family homes in Northridge, 100 homes were selected for inspection. Describe the population and sample for this problem.

1.6 A consumer group conducts crash tests of new model cars. To determine the severity of damage to 2003 Mazda 626s resulting from a 10-mph crash into a concrete wall, the research group tests six cars of this type and assesses the amount of damage. Describe the population and sample for this problem.

1.7 A building contractor has a chance to buy an odd lot of 5000 used bricks at an auction. She is interested in determining the proportion of bricks in the lot that are cracked and therefore unusable for her current project, but she does not have enough time to inspect all 5000 bricks. Instead, she checks 100 bricks to determine whether each is cracked. Describe the population and sample for this problem.

▪ 1.4 Types of Data and Some Simple Graphical Displays

Every discipline has its own particular way of using common words, and statistics is no exception. You will recognize some of the terminology from previous math and science courses, but much of the language of statistics will be new to you.

▪ Describing Data

The individuals or objects in any particular population typically possess many characteristics that might be studied. Consider a group of students currently enrolled in a statistics course. One characteristic of the students in the population is the brand of calculator owned (Casio, Hewlett-Packard, Sharp, Texas Instruments, and so on). Another characteristic is the number of textbooks purchased, and yet another is the distance from the university to each student's permanent residence. A **variable** is any characteristic whose value may change from one individual or object to another. For example, *calculator brand* is a variable, and so are *number of textbooks purchased* and *distance to the university*. **Data** result from making observations either on a single variable or simultaneously on two or more variables.

A **univariate data set** consists of observations on a single variable made on individuals in a sample or population. There are two types of univariate data sets: categorical and numerical. In the previous example, *calculator brand* is a categorical variable, because each student's response to the query, "What brand of calculator do you own?" is a category. The collection of responses from all these students forms a **categorical data set**. The other two attributes, *number of textbooks purchased* and *distance to the university*, are both numerical in nature. Determining the value of such a numerical variable (by counting or measuring) for each student results in a **numerical data set**.

▪ **Definition**

A data set consisting of observations on a single attribute is a **univariate data set**. A univariate data set is **categorical** (or **qualitative**) if the individual observations are categorical responses. A univariate data set is **numerical** (or **quantitative**) if each observation is a number.

▪ Example 1.6 Airline Safety Violations

The Federal Aviation Administration (FAA) monitors airlines and can take administrative actions for safety violations. Information about the fines assessed by the FAA appeared in the article "Just How Safe Is That Jet?" (*USA Today*, March 13, 2000). Violations that could lead to a fine were categorized as Security (S), Maintenance (M), Flight Operations (F), Hazardous Materials (H), or Other (O). Data for the variable *type of violation* for 20 administrative actions are given in the following list (these data are a subset of the data described in the article, but they are consistent with summary values given in the paper; for a description of the full data set, see Exercise 1.16):

S	S	M	H	M	O	S	M	S	S
F	S	O	M	S	M	S	M	S	M

Because *type of violation* is a categorical (nonnumerical) response, this is a categorical data set.

In Example 1.6, the data set consisted of observations on a single variable (*type of violation*), so this is univariate data. In some studies, attention focuses simultaneously on two different attributes. For example, both height (in inches) and weight (in pounds) might be recorded for each individual in a group. The resulting data set consists of pairs of numbers, such as (68, 146). This is called a **bivariate data set**. **Multivariate data** result from obtaining a category or value for each of two or more attributes (so bivariate data are a special case of multivariate data). For example, multivariate data would result from determining height, weight, pulse rate, and systolic blood pressure for each individual in a group. Example 1.7 illustrates a bivariate data set.

▪ Example 1.7 Revisiting Airline Safety Violations

The same article referenced in Example 1.6 ("Just How Safe Is That Jet?" *USA Today*, March 13, 2000) gave data on both the number of violations and the average fine per violation for the period 1985–1998 for 10 major airlines. The resulting data are given in the following table:

Airline	Number of Violations	Average Fine per Violation ($)
Alaska	258	5038.760
America West	257	3112.840
American	1745	2693.410
Continental	973	5755.396
Delta	1280	3828.125
Northwest	1097	2643.573
Southwest	535	3925.234
TWA	642	2803.738
United	1110	2612.613
US Airways	891	3479.237

Each of the variables considered here is numerical (rather than categorical) in nature. This is an example of a bivariate numerical data set.

▪ Two Types of Numerical Data

With numerical data, it is useful to make a further distinction between *discrete* and *continuous* numerical data. Visualize a number line (Figure 1.3) for locating values of the numerical variable being studied. Every possible number (2, 3.125, 8.12976, etc.) corresponds to exactly one point on the number line. Now suppose that the variable of interest is the number of cylinders of an automobile engine. The possible values of 4, 6, and 8 are identified in Figure 1.4(a) by the dots at the points marked 4, 6, and 8. These possible values are isolated from one another on the line; around any possible value, we can place an interval that is small enough that no other possible value is included in the interval. On the other hand, the line segment in Figure 1.4(b) identifies a plausible set of possible values for the time it takes a car to travel one-quarter mile. Here the possible values make up an entire interval on the number line, and no possible value is isolated from the other possible values.

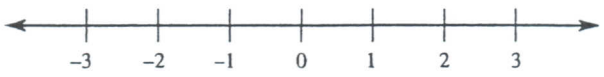


Figure 1.3 A number line.

Figure 1.4 Possible values of a variable: (a) number of cylinders; (b) quarter-mile time.

▪ Definition

Numerical data are **discrete** if the possible values are isolated points on the number line. Numerical data are **continuous** if the set of possible values forms an entire interval on the number line.

Discrete data usually arise when each observation is determined by counting (e.g., the number of classes for which a student is registered or the number of petals on a certain type of flower).

▪ Example 1.8 Calls to a Drug Abuse Hotline

The number of telephone calls per day to a drug abuse hotline is recorded for 12 days. The resulting data set is

3 0 4 3 1 0 6 2 0 0 1 2

Possible values for the variable *number of calls* are 0, 1, 2, 3, . . . ; these are isolated points on the number line, so we have a sample consisting of discrete numerical data.

The observations on the variable *number of violations* in Example 1.7 are also an example of discrete numerical data. However, in the same example, the variable *average fine per violation* could be 3000, 3000.1, 3000.125, 3000.12476, or any other value in an entire interval, so the observations on this variable provide an example of continuous data. Other examples of continuous data result from determining task completion times, body temperatures, and package weights.

In general, data are continuous when observations involve making measurements, as opposed to counting. In practice, measuring instruments do not have infinite accuracy, so possible measured values, strictly speaking, do not form a continuum on the number line. However, any number in the continuum *could* be a value of the variable. The distinction between discrete and continuous data will be important in our discussion of probability models.

▪ Frequency Distributions and Bar Charts for Categorical Data

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. When the data set is categorical, a common way to present the data is in the form of a table, called a *frequency distribution*.

A **frequency distribution for categorical data** is a table that displays the possible categories along with the associated frequencies or relative frequencies.

The **frequency** for a particular category is the number of times the category appears in the data set.

The **relative frequency** for a particular category is the fraction or proportion of the time that the category appears in the data set. It is calculated as

$$\text{relative frequency} = \frac{\text{frequency}}{\text{number of observations in the data set}}$$

When the table includes relative frequencies, it is sometimes referred to as a **relative frequency distribution**.

▪ Example 1.9 Preferred Leisure Activities

Many public health efforts are directed toward increasing levels of physical activity. The article "Physical Activity in Urban White, African American, and Mexican American Women" (*Medicine and Science in Sports and Exercise* [1997]: 1608–1614) reported on physical activity patterns in urban women. The accompanying data set gives the preferred leisure-time physical activity for each of 30 Mexican American women. The following coding is used: W = walking, T = weight training, C = cycling, G = gardening, A = aerobics.

W	T	A	W	G	T	W	W	C	W
T	W	A	T	T	W	G	W	W	C
A	W	A	W	W	W	T	W	W	T

The corresponding frequency distribution is given in Table 1.1.

Table 1.1 ▪ Frequency Distribution for Preferred Activity

Category	Frequency	Relative Frequency
Walking	15	.500 ← 15/30
Weight training	7	.233 ← 7/30
Cycling	2	.067
Gardening	2	.067
Aerobics	4	.133
	30	1.000
	Total number of observations	Should total 1, but in some cases may be slightly off due to rounding

From the frequency distribution, we can see that 15 women indicated a preference for walking and that this was by far the most popular response. Equivalently, using the relative frequencies, we can say that .5 (half or 50%) of the women preferred walking.

A frequency distribution gives a tabular display of a data set. It is also common to display categorical data graphically. A bar chart is one of the most widely used types of graphical displays for categorical data.

▪ Bar Charts

A **bar chart** is a graph of the frequency distribution of categorical data. Each category in the frequency distribution is represented by a bar or rectangle, and the picture is constructed in such a way that the *area* of each bar is proportional to the corresponding frequency or relative frequency.

▪ **Bar Charts**

When to Use

Categorical data.

How to Construct

1. Draw a horizontal line, and write the category names or labels below the line at regularly spaced intervals.
2. Draw a vertical line, and label the scale using either frequency or relative frequency.
3. Place a rectangular bar above each category label. The height is determined by the category's frequency or relative frequency, and all bars should have the same width. With the same width, both the height and the area of the bar are proportional to frequency and relative frequency.

What to Look For

Frequently and infrequently occurring categories.

▪ Example 1.10 Revisiting Preferred Leisure Activity

Example 1.9 gave data on preferred leisure-time physical activity for a sample of 30 Mexican American women. Figure 1.5 shows the bar chart corresponding to the frequency distribution constructed for these data (Table 1.1). Note that the bar chart in Figure 1.5 orders the categories alphabetically, whereas the frequency distribution listed the categories in a different order. When constructing a bar graph, the order in which the categories are listed generally does not matter. The three most common choices are (1) alphabetical ordering, (2) ordering by frequency, with the most commonly occurring category first and so on, and (3) ordering to match the order in an accompanying frequency distribution.

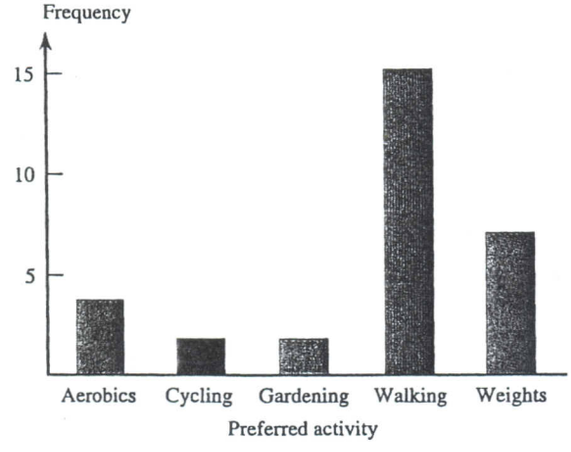


Figure 1.5 Bar chart of preferred leisure activity.

The bar chart provides a visual representation of the information in the frequency distribution. From the bar chart, it is easy to see that walking occurred most often in the data set, followed by weight training. The bar for walking is about twice

as tall (and therefore has twice the area) as the bar for weight training, because approximately twice as many women preferred walking to weight training.

▪ Example 1.11 Why Students Drop Out

The article "So Close, Yet So Far: Predictors of Attrition in College Seniors" (*Journal of College Student Development* [1998]: 343–348) examined the reasons that college seniors leave their college programs before graduating. Forty-two college seniors at a large public university who dropped out before graduation were interviewed and asked the main reason for discontinuing enrollment at the university. Data consistent with that given in the article are summarized in the following frequency distribution:

Reason for Leaving the University	Frequency
Academic problems	7
Poor advising or teaching	3
Needed a break	2
Economic reasons	11
Family responsibilities	4
To attend another school	9
Personal problems	3
Other	3

The corresponding bar chart is shown in Figure 1.6. From the bar chart, it is easy to see that more students reported leaving the university for economic reasons or to attend another school than for academic reasons.

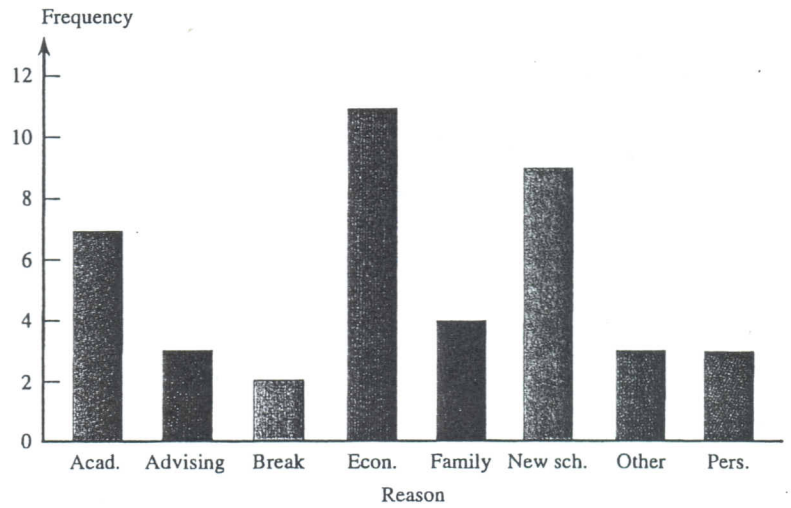


Figure 1.6 Bar chart for the data of Example 1.11, reason for leaving university.

▪ Dotplots for Numerical Data

A dotplot is a simple way to display numerical data when the data set is reasonably small. Each observation is represented by a dot above the location corresponding to its value on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence and these dots are stacked vertically.

▪ Dotplots

When to Use
Small numerical data sets.

How to Construct
1. Draw a horizontal line and mark it with an appropriate measurement scale.
2. Locate each value in the data set along the measurement scale, and represent it by a dot. If there are two or more observations with the same value, stack the dots vertically.

What to Look For
Dotplots convey information about a representative or typical value in the data set, the extent to which the data values spread out, the nature of the distribution of values along the number line, and the presence of unusual values in the data set.

▪ Example 1.12 Graduation Rates for NCAA Division I Schools in California and Texas

The Chronicle of Higher Education (Almanac Issue, August 31, 2001) reported graduation rates for NCAA Division I schools. The rates reported are the percentages of full-time freshmen in fall 1993 who had earned a bachelor's degree by August 1999. Data from the two largest states (California, with 20 Division I schools; and Texas, with 19 Division I schools) are given in the following list:

California:	64	41	44	31	37	73	72	68	35	37
	81	90	82	74	79	67	66	66	70	63
Texas:	67	21	32	88	35	71	39	35	71	63
	12	46	35	39	28	65	25	24	22	

MINITAB, a computer software package for statistical analysis, was used to construct a dotplot of the 39 graduation rates. This dotplot is given in Figure 1.7. From the dotplot, we can see that graduation rates varied a great deal from school to school and that the graduation rates seem to form two distinguishable groups of about the same size — one group with higher graduation rates and one with lower graduation rates.

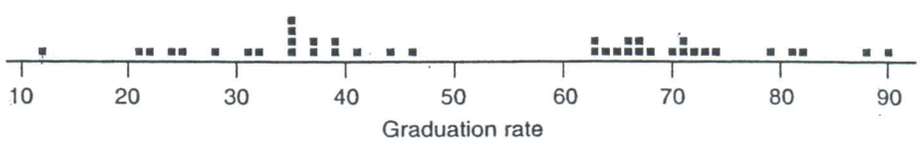


Figure 1.7 Dotplot of graduation rates.

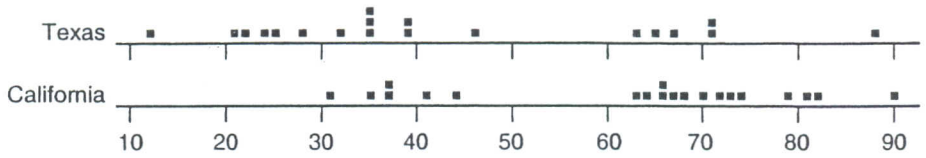


Figure 1.8 Dotplot of graduation rates for California and Texas.

Figure 1.8 shows separate dotplots for the California and Texas schools. The dotplots are drawn using the same scale to facilitate comparisons. From the two plots, we can see that, although both states have a high group and a low group, there are only six schools in the low group for California and only six schools in the high group for Texas.

▪ Exercises 1.8–1.19

1.8 Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Number of students in a class of 35 who turn in a term paper before the due date

b. Gender of the next baby born at a particular hospital

c. Amount of fluid (in ounces) dispensed by a machine used to fill bottles with soda pop

d. Thickness of the gelatin coating of a vitamin E capsule

e. Birth order classification (only child, firstborn, middle child, lastborn) of a math major

1.9 Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

a. Brand of computer purchased by a customer

b. State of birth for someone born in the United States

c. Price of a textbook

d. Concentration of a contaminant (micrograms per cubic centimeter) in a water sample

e. Zip code (Think carefully about this one.)

f. Actual weight of coffee in a 1-lb can

1.10 For the following numerical attributes, state whether each is discrete or continuous.

a. The number of insufficient-funds checks received by a grocery store during a given month

b. The amount by which a 1-lb package of ground beef decreases in weight (because of moisture loss) before purchase

c. The number of New York Yankees during a given year who will not play for the Yankees the next year

d. The number of students in a class of 35 who have purchased a used copy of the textbook

e. The length of a 1-year-old rattlesnake

f. The altitude of a location in California selected randomly by throwing a dart at a map of the state

g. The distance from the left edge at which a 12-in. plastic ruler snaps when bent sufficiently to break

h. The price per gallon paid by the next customer to buy gas at a particular station

1.11 For each of the following situations, give a set of possible data values that might arise from making the observations described.

a. The manufacturer for each of the next 10 automobiles to pass through a given intersection is noted.

b. The grade point average for each of the 15 seniors in a statistics class is determined.

c. The number of gas pumps in use at each of 20 gas stations at a particular time is determined.

d. The actual net weight of each of 12 bags of fertilizer having a labeled weight of 50 lb is determined.

e. Fifteen different radio stations are monitored during a 1-hr period, and the amount of time devoted to commercials is determined for each.

1.12 *Spider-Man* and *Star Wars: Episode II* were the top moneymakers among the summer 2002 movie releases. Box office totals for the top summer films in 2002 are given in the following table (*USA Today*, September 3, 2002):

Film	Box Office (millions of dollars)
Spider-Man	403.7
Star Wars: Episode II	300.1
Austin Powers in Goldmember	203.5
Signs	195.1
Men in Black II	189.7
Scooby-Doo	151.6
Lilo & Stitch	141.8
Minority Report	130.1
Mr. Deeds	124.0
XXX	123.9
The Sum of All Fears	118.5
The Bourne Identity	118.1
Road to Perdition	99.1
My Big Fat Greek Wedding	82.3
Spirit: Stallion of the Cimarron	73.2
Spy Kids 2: Island of Lost Dreams	69.1
Divine Secrets of the Ya-Ya Sisterhood	68.8
Insomnia	67.0
Stuart Little 2	61.9
Unfaithful	52.8

Use a dotplot to display these data. Write a few sentences commenting on the notable features of the dotplot.

1.13 Water quality ratings of 36 Southern California beaches were given in the article "How Safe Is the Surf?" (*Los Angeles Times*, October 26, 2002). The ratings, which ranged from A+ to F and which reflect the risk of getting sick from swimming at a particular beach, are given in the following list:

A+	A+	A+	F	A+	A+
A	B	A	C	C	A−
A	F	B	A+	C	A
D	F	A+	D	A+	A
D	A	A	D	A+	A+
A	A	C	F−	B	A+

a. Summarize the given ratings by constructing a relative frequency distribution and a bar chart. Comment on the interesting features of your bar chart.
b. Would it be appropriate to construct a dotplot for these data? Why or why not?

1.14 Many adolescent boys aspire to be professional athletes. The paper "Why Adolescent Boys Dream of Becoming Professional Athletes" (*Psychological Reports* [1999]: 1075–1085) examined some of the reasons. Each boy in a sample of teenage boys was asked the following question: "Previous studies have shown that more teenage boys say that they are considering becoming professional athletes than any other occupation. In your opinion, why do these boys want to become professional athletes?" The resulting data are shown in the following table:

Response	Frequency
Fame and celebrity	94
Money	56
Attract women	29
Like sports	27
Easy life	24
Don't need an education	19
Other	19

Construct a bar chart to display these data.

1.15 The article "Knee Injuries in Women Collegiate Rugby Players" (*American Journal of Sports Medicine* [1997]: 360–362) reported the following data on type of injury sustained by 13 female rugby players (MCL, ACL, and the meniscus are different sites in the knee, and the patella is the kneecap). These data are a subset of the data given in the article:

meniscus tear	patella dislocation	MCL tear
meniscus tear	ACL tear	meniscus tear
meniscus tear	MCL tear	meniscus tear
ACL tear	patella	MCL tear
MCL tear	dislocation	

Summarize these data in a frequency distribution, and then construct a bar chart for this data set. Write a sentence or two describing the relative occurrence of the different types of injury.

1.16 The article "Just How Safe Is That Jet?" (*USA Today*, March 13, 2000) gave the following relative frequency distribution that summarized data on the type of violation for fines imposed on airlines by the Federal Aviation Administration:

Type of Violation	Relative Frequency
Security	.43
Maintenance	.39
Flight operations	.06
Hazardous materials	.03
Other	.09

Use this information to construct a bar chart for type of violation, and then write a sentence or two commenting on the relative occurrence of the various types of violation.

1.17 The article "Americans Drowsy on the Job and The Road" (Associated Press, March 28, 2001) summarized data from the 2001 Sleep in America poll. Each individual in a sample of 1004 adults was asked

questions about his or her sleep habits. The article states that "40 percent of those surveyed say they get sleepy on the job and their work suffers at least a few days each month, while 22 percent said the problems occur a few days each week. And 7 percent say sleepiness on the job is a daily occurrence." Assuming that everyone else reported that sleepiness on the job was not a problem, summarize the given information by constructing a relative frequency bar chart.

1.18 "Ozzie and Harriet Don't Live Here Anymore" (*San Luis Obispo Tribune*, February 26, 2002) is the title of an article that looked at the changing makeup of America's suburbs. The article states that nonfamily households (e.g., homes headed by a single professional or an elderly widow) now outnumber married couples with children in suburbs of the nation's largest metropolitan areas. The article goes on to state:

> In the nation's 102 largest metropolitan areas, "nonfamilies" comprised 29 percent of households in 2000, up from 27 percent in 1990. While the number of married-with-children homes grew too, the share did not keep pace. It declined from 28 percent to 27 percent. Married couples without children at home live in another 29 percent of suburban households. The remaining 15 percent are single-parent homes.

Use the given information on type of household in 2000 to construct a frequency distribution and a bar chart. (Be careful to extract the 2000 percentages from the given information).

1.19 Each year *U.S. News and World Report* publishes a ranking of U.S. business schools. The following data give the acceptance rates (percentage of applicants admitted) for the best 25 programs in the most recent survey:

16.3	12.0	25.1	20.3	31.9	20.7	30.1	19.5	36.2
46.9	25.8	36.7	33.8	24.2	21.5	35.1	37.6	23.9
17.0	38.4	31.2	43.8	28.9	31.4	48.9		

Construct a dotplot, and comment on the interesting features of the plot.

▪ Activity 1.1: Head Sizes: Understanding Variability

Materials needed: Each team will need a measuring tape.

For this activity, you will work in teams of 6 to 10 people.

1. Designate a team leader for your team by choosing the person on your team who celebrated his or her last birthday most recently.

2. The team leader should measure and record the head size (measured as the circumference at the widest part of the forehead) of each of the other members of his or her team.

3. Record the head sizes for the individuals on your team as measured by the team leader.

4. Next, each individual on the team should measure the head size of the team leader. Do not share your measurement with the other team members until all team members have measured the team leader's head size.

5. After all team members have measured the team leader's head, record the different team leader head size measurements obtained by the individuals on your team.

6. Using the data from Step 3, construct a dotplot of the team leader's measurements of team head sizes. Then, using the same scale, construct a separate dotplot of the different measurements of the team leader's head size (from Step 5).

Now use the available information to answer the following questions:

7. Do you think the team leader's head size changed in between measurements? If not, explain why the measurements of the team leader's head size are not all the same.

8. Which data set was more variable — head size measurements of the different individuals on your team or the different measurements of the team leader's head size? Explain the basis for your choice.

9. Consider the following scheme (you don't actually have to carry this out): Suppose that a group of 10 people measured head sizes by first assigning each person in the group a number between 1 and 10. Then person 1 measured person 2's head size, person 2 measured person 3's head size, and so on, with person 10 finally measuring person 1's head size. Do you think that the resulting head size measurements would be more variable, less variable, or show about the same amount of variability as a set of 10 measurements resulting from a single individual measuring the head size of all 10 people in the group? Explain.

▪ Activity 1.2: Estimating Sizes

1. Construct an activity sheet that consists of a table that has 6 columns and 10 rows. Label the columns of the table with the following six headings: (1) Shape, (2) Estimated Size, (3) Actual Size, (4) Difference (Est. − Actual), (5) Absolute Difference, and (6) Squared Difference. Enter the numbers from 1 to 10 in the "Shape" column.

2. Next you will be visually estimating the sizes of the shapes in Figure 1.9. Size will be described as the number of squares of this size

that would fit in the shape. For example, the shape

would be size 3, as illustrated by

You should now quickly *visually* estimate the sizes of the shapes in Figure 1.9. *Do not* draw on the figure — these are to be quick visual estimates. Record your estimates in the "Estimated Size" column of the activity sheet.

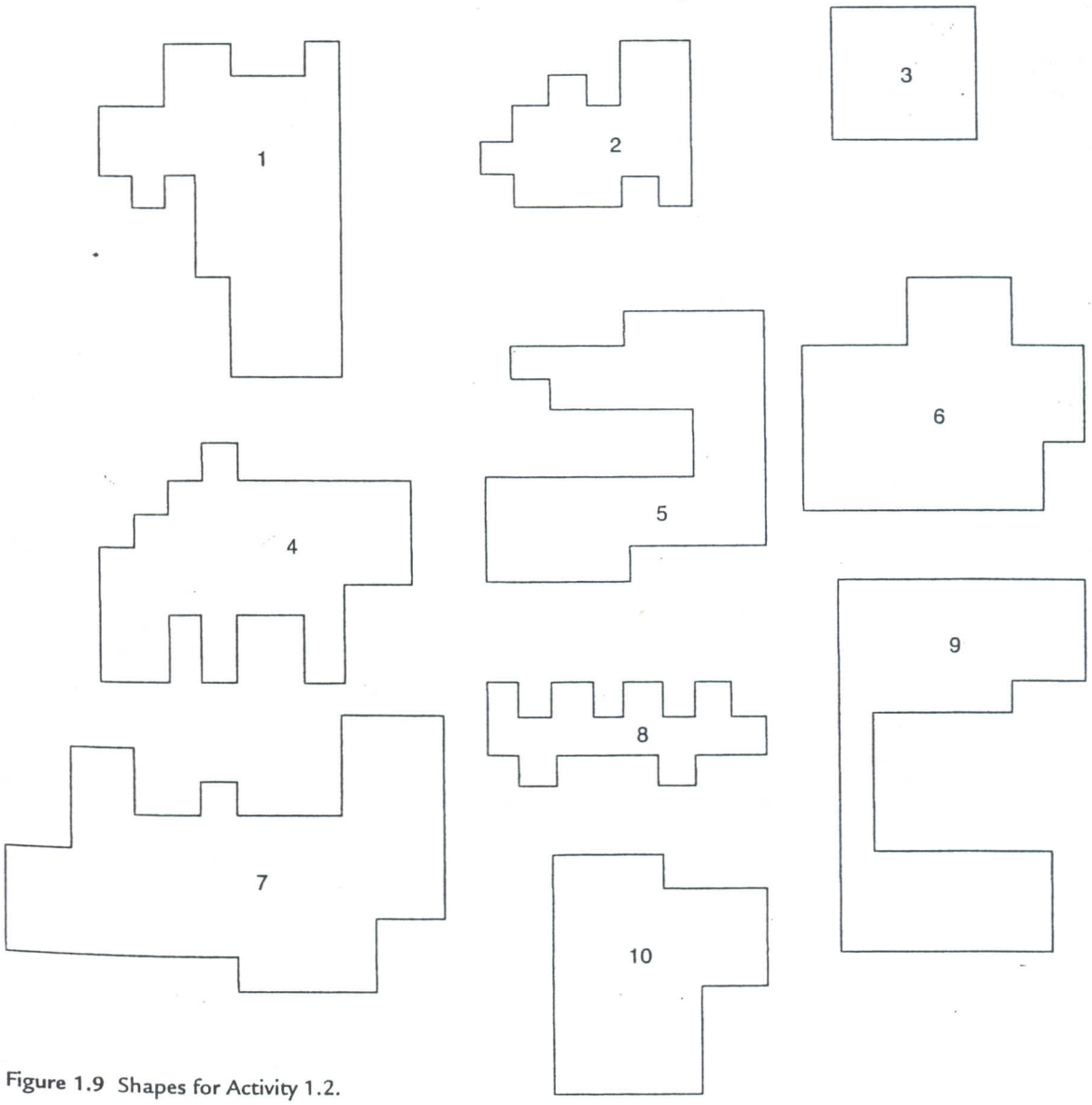


Figure 1.9 Shapes for Activity 1.2.

3. Your instructor will provide the actual sizes for the 10 shapes, which should be entered into the "Actual Size" column of the activity sheet. Now complete the "Difference" column by subtracting the actual value from your estimate for each of the 10 shapes.

4. What would cause a difference to be negative? positive?

5. Would the sum of the differences tell you if the estimates and actual values were in close agreement? Does a sum of 0 for the differences indicate that all the estimates were equal to the actual value? Explain.

6. Compare your estimates with those of another person in the class by comparing the sum of the absolute values of the differences between estimates and corresponding actual values. Who was better at estimating shape sizes? How can you tell?

7. Us... the sq... shape... $(-3)^2$ ... ference... your sha...

8. For this... dents from... form a new... of the size esti... viduals in your... more accura... were? How can...

9. Does your ans... plain why or why no...

▪ Summary of Key Concepts and Formulas

Term or Formula	Comment
Descriptive statistics	Numerical, graphical, and tabular methods for ... rizing data.
Population	The entire collection of individuals or measurem... formation is desired.
Sample	A part of the population selected for study.
Categorical data	Individual observations are categorical responses (nonnum...
Numerical data	Individual observations are numerical (quantitative) in na...
Discrete numerical data	Possible values are isolated points along the number line.
Continuous numerical data	Possible values form an entire interval along the number line.
Bivariate and multivariate data	Each observation consists of two (bivariate) or more (multivaria... responses or values.
Frequency distribution for categorical data	A table that displays frequencies, and sometimes relative frequen... for each of the possible values of a categorical variable.
Bar chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a bar, and the area of the bar is proportiona... to the corresponding frequency or relative frequency.
Dotplot	A picture of numerical data in which each observation is represented by a dot on or above a horizontal measurement scale.

▪ Supplementary Exercises 1.20–1.24

1.20 The paper "Profile of Sport/Leisure Injuries Treated at Emergency Rooms of Urban Hospitals" (*Canadian Journal of Sports Science* [1991]: 99–102) classified noncontact sports injuries by sport, resulting in the following table:

Sport	Number of Sport Injuries
Touch football	38
Soccer	24
Basketball	19
Baseball/softball	11
Jogging/running	11
Bicycling	11
Volleyball	17
Other	47

Calculate relative frequencies and draw the corresponding bar chart.

1.21 Nonresponse is a common problem facing researchers who rely on mail questionnaires. In the paper "Reasons for Nonresponse on the Physicians' Practice Survey" (*1980 Proceedings of the Section on Social Statistics* [Alexandria, VA: American Statistical Association, 1980]: 202), 811 doctors who did not respond to the AMA Survey of Physicians were contacted about the reason for their nonparticipation. The results are summarized in the accompanying relative frequency distribution. Draw the corresponding bar chart.

Reason	Relative Frequency
1. No time to participate	.264
2. Not interested	.300
3. Don't like surveys in general	.145
4. Don't like this particular survey	.025
5. Hostility toward the government	.054
6. Desire to protect privacy	.056
7. Other reason for refusal	.053
8. No reason given	.103

1.22 The paper "Fire Doors: A Potential Weak Link in the Protection Chain" (*Fire Technology* [1992]: 177–179) reported on a survey of companies that use fire doors to subdivide major plant facilities in case of uncontrolled fires. Four types of doors were in use: rolling steel (R), single metal-clad sliding (M), swinging with door closer (C), and single swinging (S). Suppose that the data were as shown in the following list (percentages agree with those given in the paper):

R	M	M	S	R	R	C	R	M	R
M	M	M	C	S	R	R	C	C	M
R	R	C	S	R	C	C	M	S	M
M	M	R	S	R	R	C	C	S	R
M	C	R	M	C	M	R	C	M	S

Determine the frequencies and the relative frequencies for the four categories, and display them in a relative frequency table. Construct a bar graph for this data set, and write a sentence or two commenting on the interesting features of this display.

1.23 The article "Can We Really Walk Straight?" (*American Journal of Physical Anthropology* [1992]: 19–27) reported on an experiment in which each of 20 healthy men was asked to walk as straight as possible to a target 60 m away at normal speed. Consider the following observations on cadence (number of strides per second):

0.95	0.85	0.92	1.95	0.93	1.86	1.00
0.92	0.85	0.81	0.78	0.93	0.93	1.05
0.93	1.06	1.06	0.96	0.81	0.96	

Construct a dotplot for the cadence data. Do any data values stand out as being unusual?

1.24 A sample of 50 individuals who recently joined a certain travel club yielded the responses on occupation shown in the following table (C = clerical, M = manager/executive, P = professional, R = retired, S = sales, T = skilled tradesperson, O = other):

P	R	R	M	S	R	P	R	C	P
M	R	T	O	R	S	R	S	P	M
R	P	C	R	R	S	T	P	P	C
M	S	R	R	P	R	S	R	R	P
M	P	R	R	S	C	P	R	S	M

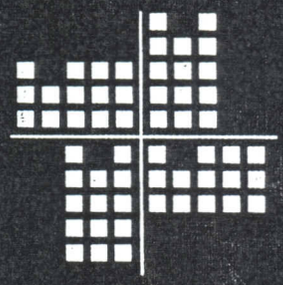
Summarize these data using a graphical display.

▪ References

Moore, David. *Statistics: Concepts and Controversies*, 4th ed. New York: W. H. Freeman, 1997. (A nice, informal survey of statistical concepts and reasoning.)

Tanur, Judith, ed. *Statistics: A Guide to the Unknown*. Belmont, CA: Duxbury Press, 1989. (Short, nontechnical articles by a number of well-known statisticians and users of statistics on the application of statistics in various disciplines and subject areas.)

Utts, Jessica. *Seeing Through Statistics*. Belmont, CA: Duxbury Press, 1999. (A nice introduction to the fundamental ideas of statistical reasoning.)



2 ▪ The Data Analysis Process and Collecting Data Sensibly

The data analysis process can be viewed as a sequence of steps that lead from planning to data collection to informed conclusions based on the resulting data. It should come as no surprise that the correctness of such conclusions depends on the quality of the information on which they are based. The data collection step is critical to obtaining reliable information; both the type of analysis that is appropriate and the conclusions that can be drawn depend on how the data are collected. In this chapter, we first describe the data analysis process in some detail and then focus on two widely used methods of data collection: sampling and experimentation.

▪ 2.1 The Data Analysis Process

Statistics involves the collection and analysis of data. Both tasks are critical. Raw data without analysis are of little value, and even a sophisticated analysis cannot extract meaningful information from data that were not collected in a sensible way.

▪ Planning and Conducting a Study

Scientific studies are undertaken to answer questions about our world. Is a new flu vaccine effective in preventing illness? Is the use of bicycle helmets on the rise? Are injuries that result from bicycle accidents less severe for riders who wear helmets than for those who do not? How many credit cards do college students have? Do engineering students or psychology students pay more for textbooks? Data collection and analysis allow researchers to answer such questions.

The data analysis process can be organized into the following six steps:

1. **Understanding the nature of the problem.** Effective data analysis requires an understanding of the research problem. We must know the goal of the research and what questions we hope to answer. It is important to have a clear direction before gathering data to avoid being unable to answer the questions of interest using the data collected.

2. **Deciding what to measure and how to measure it:** The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious (e.g., in a study of the relationship between the weight of a Division I football player and position played, you would need to collect data on player weight and position), but in other cases the choice of information is not as straightforward (e.g., in a study of the relationship between preferred learning style and intelligence, how would you define learning style and measure it and what measure of intelligence would you use?). It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.

Simple Random Sample (SRS)
Systematic Sampling
Stratified Sampling
Cluster Sampling

3. **Data collection.** The data collection step is crucial. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. Even if a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood and judged to be acceptable. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the subsequent conclusions that can be drawn depend on how the data are collected.
4. **Data summarization and preliminary analysis.** After the data are collected, the next step usually involves a preliminary analysis that includes summarizing the data graphically and numerically. This initial analysis provides insight into important characteristics of the data and can provide guidance in selecting appropriate methods for further analysis.
5. **Formal data analysis.** The data analysis step requires the researcher to select and apply the appropriate statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.
6. **Interpretation of results.** Several questions should be addressed in this final step — for example, What conclusions can be drawn from the analysis? How do the results of the analysis inform us about the stated research problem or question? and How can our results guide future research? The interpretation step often leads to the formulation of new research questions, which, in turn, leads back to the first step. In this way, good data analysis is often an iterative process.

Example 2.1 illustrates the steps in the data analysis process.

▪ Example 2.1 A Proposed New Treatment for Alzheimer's Disease

The article "Brain Shunt Tested to Treat Alzheimer's" (*San Francisco Chronicle*, October 23, 2002) summarizes the findings of a study that appeared in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach to treating Alzheimer's disease results in improved memory functioning. The surgical procedure involves implanting a thin tube, called a shunt, that is designed to drain toxins from the fluid-filled space that cushions the brain. Eleven patients had shunts implanted and were followed for a year, receiving quarterly tests of memory function. Another group of Alzheimer's patients was used as a comparison group. The patients in the comparison group received the standard care for Alzheimer's disease. After analyzing the data from this study, the investigators concluded that the "results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily

declined. However, the study was too small to produce conclusive statistical evidence." Based on these results, a much larger 18-month study is being planned. That study will include 256 patients at 25 medical centers around the country; the results are expected in 2005.

This study illustrates the nature of the data analysis process. A clearly defined research question and an appropriate choice for how to measure the variable of interest (the cognitive tests used to measure memory function) preceded the data collection. Assuming that a reasonable method was used to collect the data (we will see how this can be evaluated in Sections 2.4 and 2.5) and that appropriate methods of analysis were employed, the investigators reached the conclusion that the surgical procedure showed promise. However, they recognized the limitations of the study, especially those resulting from the small number of patients in the group that received surgical treatment, which in turn led to the design of a larger, more sophisticated study. As is often the case, the data analysis cycle led to further research, and the process began anew.

▪ Evaluating a Research Study

The six data analysis steps can also be used as a guide for evaluating published research studies. The following questions should be addressed as part of a study evaluation:

- What were the researchers trying to learn? What questions motivated their research?
- Was relevant information collected? Were the right things measured?
- Were the data collected in a sensible way?
- Were the data summarized in an appropriate way?
- Was an appropriate method of analysis used, given the type of data and how the data were collected?
- Are the conclusions drawn by the researchers supported by the data analysis?

Example 2.2 illustrates how these questions can guide an evaluation of a research study.

▪ Example 2.2 Spray Away the Flu

The newspaper article "Spray Away Flu" (*Omaha World-Herald*, June 8, 1998) reported on a study of the effectiveness of a new flu vaccine that is administered by nasal spray rather than by injection. The article states that the "researchers gave the spray to 1070 healthy children, 15 months to 6 years old, before the flu season two winters ago. One percent developed confirmed influenza, compared with 18 percent of the 532 children who received a placebo. And only one vaccinated child developed an ear infection after coming down with influenza. . . . Typically 30 percent to 40 percent of children with influenza later develop an ear infection." The researchers concluded that the nasal flu vaccine was effective in reducing the incidence of flu and also in reducing the number of children with flu who subsequently develop ear infections.

In this study, the researchers were trying to find out whether the nasal flu vaccine was effective in reducing the number of flu cases and in reducing the number

2. **Deciding what to measure and how to measure it.** The next step in the process is deciding what information is needed to answer the questions of interest. In some cases, the choice is obvious (e.g., in a study of the relationship between the weight of a Division I football player and position played, you would need to collect data on player weight and position), but in other cases the choice of information is not as straightforward (e.g., in a study of the relationship between preferred learning style and intelligence, how would you define learning style and measure it and what measure of intelligence would you use?). It is important to carefully define the variables to be studied and to develop appropriate methods for determining their values.

Simple Random Sample (SRS)
Systematic Sampling
Stratified Sampling
Cluster Sampling

3. **Data collection.** The data collection step is crucial. The researcher must first decide whether an existing data source is adequate or whether new data must be collected. Even if a decision is made to use existing data, it is important to understand how the data were collected and for what purpose, so that any resulting limitations are also fully understood and judged to be acceptable. If new data are to be collected, a careful plan must be developed, because the type of analysis that is appropriate and the subsequent conclusions that can be drawn depend on how the data are collected.
4. **Data summarization and preliminary analysis.** After the data are collected, the next step usually involves a preliminary analysis that includes summarizing the data graphically and numerically. This initial analysis provides insight into important characteristics of the data and can provide guidance in selecting appropriate methods for further analysis.
5. **Formal data analysis.** The data analysis step requires the researcher to select and apply the appropriate statistical methods. Much of this textbook is devoted to methods that can be used to carry out this step.
6. **Interpretation of results.** Several questions should be addressed in this final step — for example, What conclusions can be drawn from the analysis? How do the results of the analysis inform us about the stated research problem or question? and How can our results guide future research? The interpretation step often leads to the formulation of new research questions, which, in turn, leads back to the first step. In this way, good data analysis is often an iterative process.

Example 2.1 illustrates the steps in the data analysis process.

▪ Example 2.1 A Proposed New Treatment for Alzheimer's Disease

The article "Brain Shunt Tested to Treat Alzheimer's" (*San Francisco Chronicle*, October 23, 2002) summarizes the findings of a study that appeared in the journal *Neurology*. Doctors at Stanford Medical Center were interested in determining whether a new surgical approach to treating Alzheimer's disease results in improved memory functioning. The surgical procedure involves implanting a thin tube, called a shunt, that is designed to drain toxins from the fluid-filled space that cushions the brain. Eleven patients had shunts implanted and were followed for a year, receiving quarterly tests of memory function. Another group of Alzheimer's patients was used as a comparison group. The patients in the comparison group received the standard care for Alzheimer's disease. After analyzing the data from this study, the investigators concluded that the "results suggested the treated patients essentially held their own in the cognitive tests while the patients in the control group steadily

of ear infections in children who did get the flu. The researchers recorded whether a child received the nasal vaccine or a placebo. (A **placebo** is a treatment that has the appearance of the treatment of interest but contains no active ingredients). Whether or not the child developed the flu and a subsequent ear infection was also recorded. These are appropriate measurements to make in order to answer the research question of interest. We typically cannot tell much about the data collection process from a newspaper article. As we will see in Section 2.4, to fully evaluate this study, we would also want to know how the participating children were selected, how it was determined that a particular child received the vaccine or the placebo, and how the subsequent diagnoses of flu and ear infection were made.

We will also have to delay discussion of the data analysis and the appropriateness of the conclusions because we do not yet have the necessary tools to evaluate these aspects of the study.

Other interesting examples of statistical studies can be found in Judith Tanur's *Statistics: A Guide to the Unknown* and in Roger Hock's *Forty Studies That Changed Psychology: Exploration into the History of Psychological Research* (the complete references for these two books can be found at the end of this chapter).

▪ 2.2 Sampling

There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**). The most common reason is limited resources; restrictions on available time or money usually prohibit observation of an entire population. Sometimes the process of measuring the characteristics of interest is destructive, as with measuring the breaking strength of soda bottles, the lifetime of flashlight batteries, or the sugar content of oranges, and it would be foolish to study the entire population.

Many studies are conducted in order to generalize from a sample to the corresponding population. As a result, it is important that the sample be representative of the population. To be reasonably sure of this, the researcher must carefully consider the way in which the sample is selected. It is sometimes tempting to take the easy way out and gather data in a haphazard way; but if a sample is chosen on the basis of convenience alone, it becomes impossible to interpret the resulting data with confidence. For example, it might be easy to use the students in your statistics class as a sample of students at your university. However, not all majors include a statistics course in their curriculum, and most students take statistics in their sophomore or junior year. The difficulty is that it is not clear whether or how these factors (and others that we might not be aware of) affect inferences based on information from such a sample. *There is no way to tell just by looking at a sample whether it is representative of the population from which it was drawn. Our only assurance comes from the method used to select the sample.*

▪ Bias in Sampling

Bias in sampling results in the tendency for a sample to differ from the corresponding population in some systematic way. Bias can result from the way in which the

sample is selected or from the way in which information is obtained once the sample has been chosen. The most common types of bias encountered in sampling situations are selection bias, measurement or response bias, and nonresponse bias.

Selection bias is introduced when the way the sample is selected systematically excludes some part of the population of interest. This problem is sometimes also called undercoverage. For example, a researcher may wish to generalize from the results of a study to the population consisting of all residents of a particular city, but the method of selecting individuals may exclude the homeless or those without telephones. If those who are excluded from the sampling process differ in some systematic way from those who are included, the sample is guaranteed to be unrepresentative of the population. If this difference between the included and the excluded occurs on a variable that is important to the study, conclusions based on the sample data may not be valid. Selection bias also occurs if only volunteers or self-selected individuals are used in a study, because self-selected individuals (e.g., those who choose to participate in a call-in telephone poll) may well differ from those who choose not to participate.

Measurement or **response bias** occurs when the method of observation tends to produce values that systematically differ from the true value in some way. This might happen if an improperly calibrated scale is used to weigh items or if questions on a survey are worded in a way that tends to influence the response. For example, a May 1994 Gallup survey sponsored by the American Paper Institute (*Wall Street Journal*, May 17, 1994) included the following question: "It is estimated that disposable diapers account for less than 2 percent of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard waste are estimated to account for about 21 percent of trash in landfills. Given this, in your opinion, would it be fair to tax or ban disposable diapers?" It is likely that the wording of this question prompted people to respond in a particular way.

Other things that might contribute to response bias are the appearance or behavior of the person asking the question, the group or organization conducting the study, and the tendency for people to not be completely honest when asked about illegal behavior or unpopular beliefs.

Although the terms *measurement bias* and *response bias* are often used interchangeably, the term *measurement bias* is usually used to describe systematic deviation from the true value as a result of a faulty measurement instrument (as with the improperly calibrated scale).

Nonresponse bias occurs when responses are not actually obtained from all individuals selected for inclusion in the sample. As with selection bias, nonresponse bias can distort results if those who respond differ in important ways from those who do not respond. Although some level of nonresponse is unavoidable in most surveys, the biasing effect on the resulting sample is lowest when the response rate is high. To minimize nonresponse bias, it is critical that a serious effort be made to follow up with individuals who do not respond to an initial request for information.

The nonresponse rate for surveys or opinion polls varies dramatically, depending on how the data are collected. Surveys are commonly conducted by mail, by phone, or by personal interview. Mail surveys are inexpensive but often have high nonresponse rates. Telephone surveys can also be inexpensive and can be implemented quickly, but they work well only for short surveys and they can also have high nonresponse rates. Personal interviews are generally expensive but tend to have better response rates. Some of the many challenges of conducting surveys are discussed in Section 2.6.

▪ Types of Bias

Selection Bias

Tendency for samples to differ from the corresponding population as a result of systematic exclusion of some part of the population.

Measurement or Response Bias

Tendency for samples to differ from the corresponding population because the method of observation tends to produce values that differ from the true value.

Nonresponse Bias

Tendency for samples to differ from the corresponding population because data are not obtained from all individuals selected for inclusion in the sample.

It is important to note that bias is introduced by the way in which a sample is selected or by the way in which the data are collected from the sample. Increasing the size of the sample, although possibly desirable for other reasons, does nothing to reduce bias. A good discussion of the types of bias appears in the sampling book by Lohr listed in the references at the end of this chapter.

▪ Random Sampling

Most of the inferential methods introduced in this text are based on the idea of random selection. The most straightforward of these methods is called simple random sampling. A **simple random sample** is a sample chosen using a method that ensures that each different possible sample of the desired size has an equal chance of being the one chosen. For example, suppose that we want a simple random sample of 10 employees chosen from all those who work at a large design firm. For the sample to be a simple random sample, each of the many different subsets of 10 employees must be equally likely to be the one selected. A sample taken from only full-time employees would not be a simple random sample of *all* employees, because someone who works part-time is still considered an employee but has no chance of being selected. Although a simple random sample may, by chance, include only full-time employees, it must be selected in such a way that each possible sample, and therefore *every* employee, has the same chance of inclusion in the sample. It is the selection process, not the final sample, that determines whether the sample is a simple random sample.

The letter n is used to denote sample size; it is the number of individuals or objects in the sample. For the design firm scenario of the previous paragraph, $n = 10$.

▪ Definition

A simple random sample of size n is a sample that is selected from a population in a way that ensures that every different possible sample of the desired size has the same chance of being selected.

The definition of a simple random sample implies that every individual member of the population has an equal chance of being selected. *However, the fact that every individual has an equal chance of selection, by itself, is not enough to guarantee that the sample is a simple random sample.* For example, suppose that a class is

made up of 100 students, 60 of whom are female. A researcher decides to select 6 of the female students by writing all 60 names on slips of paper, mixing the slips, and then picking 6. She then selects four male students from the class using a similar procedure. Even though every student in the class has an equal chance of being included in the sample (6 of 60 females are selected and 4 of 40 males are chosen), the resulting sample is *not* a simple random sample because not all different possible samples of 10 students from the class have the same chance of selection. Many possible samples of ten students — for example, a sample of seven females and three males or a sample of all females — have no chance of being selected. The sample selection method described here is not necessarily a bad choice (in fact, it is an example of a stratified sample, to be discussed in more detail shortly), but it is not a simple random sample; this must be considered when a method is chosen for analyzing data from such a sample.

▪ **Selecting a Simple Random Sample** A number of different methods can be used to select a simple random sample. One way is to put the name or number of each member of the population on a different slip of paper; the slips are otherwise identical. The process of thoroughly mixing the slips and then selecting n slips one by one yields a random sample of size n. This method is easy to understand, but it has obvious drawbacks. The mixing must be adequate, and producing the necessary slips of paper can be extremely tedious, even for populations of moderate size.

A commonly used method for selecting a random sample is to first create a list, called a **sampling frame**, of the objects or individuals in the population. Each item on the list can then be identified by a number, and a table of random digits or a random number generator can be used to select the sample. A random number generator is a procedure that produces a sequence of numbers that satisfies all reasonable properties associated with the notion of randomness. Most statistics software packages include a random number generator, as do many calculators. A small table of random digits can be found in Appendix Table 1.

When selecting a random sample, researchers can choose to do the sampling with or without replacement. **Sampling with replacement** means that after each successive item is selected for the sample, the item is "replaced" back into the population and may therefore be selected again at a later stage. Thus, sampling with replacement allows for the possibility of having the same item or individual appear more than once in the sample. In practice, sampling with replacement is rarely used. Instead, the more common method is to not allow the same item to be included in the sample more than once. After being included in the sample, an individual or object would not be considered for further selection. Sampling in this manner is called **sampling without replacement**.

Sampling Without Replacement
Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n distinct individuals from the population.

Sampling with Replacement
After an individual from the population is selected for inclusion in the sample and the corresponding data are recorded, the individual is placed back in the population and can be selected again in the sampling process. A sample selected with replacement might include any particular individual from the population more than once.

Although these two forms of sampling are different, when the sample size n is small relative to the population size, as is often the case, there is little practical difference between them. In practice, the two can be viewed as equivalent if the sample size is at most 5% of the population size.

▪ Example 2.3 Selecting a Random Sample of Glass Soda Bottles

Breaking strength is an important characteristic of glass soda bottles. If the strength is too low, a bottle may burst — not a desirable outcome. Suppose that we want to measure the breaking strength of each bottle in a random sample of size $n = 3$ selected from four crates containing a total of 100 bottles (the population). Each crate contains five rows of five bottles each. We can identify each bottle with a number from 1 to 100 by numbering across the rows, starting with the top row of crate 1, as pictured:

Crate 1

1	2	3	4	5
6	...			

Crate 2

26	27	28	...	

...

Crate 4

76	77	...		
				100

Using a random number generator from a calculator or statistical software package, we could generate three random numbers between 1 and 100 to determine which bottles would be included in our sample. This might result in bottles 15 (row 3 column 5 of crate 1), 89 (row 3 column 4 of crate 4), and 60 (row 2 column 5 of crate 3) being selected. (Alternatively, we could write the numbers from 1 to 100 on slips of paper, place them in a container, mix them well, and then select three.)

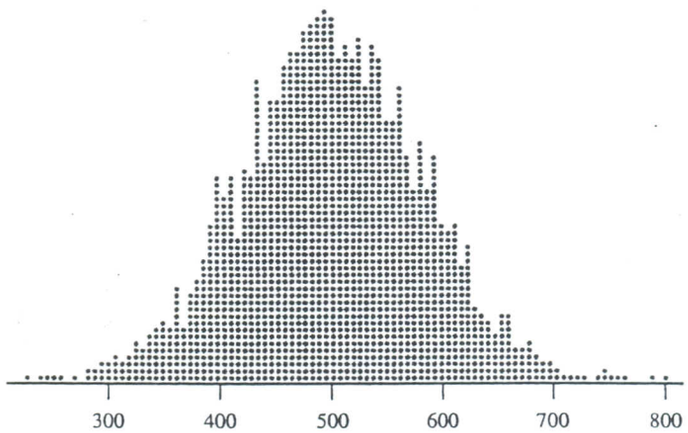
The goal of random sampling is to produce a sample that is likely to be representative of the population. Although random sampling does not *guarantee* that the sample will be representative, probability-based methods can be used to assess the risk of an unrepresentative sample. It is the ability to quantify this risk that allows us to generalize with confidence from a random sample to the corresponding population.

▪ A Note Concerning Sample Size

It is a common misconception that if the size of a sample is relatively small compared to the population size, the sample can't possibly accurately reflect the population. Critics of polls often make statements such as, "There are 14.6 million registered voters in California. How can a sample of 1000 registered voters possibly reflect public opinion when only about 1 in every 14,000 people are included in the sample?" These critics do not understand the power of random selection!

Consider a population consisting of 5000 applicants to a state university, and suppose that we are interested in math SAT scores for this population. A dotplot

of the values in this population is shown in Figure 2.1(a). Figure 2.1(b) shows dotplots of the math SAT scores for individuals in four different random samples from the population, ranging in sample size from $n = 50$ to $n = 1000$. Notice that the samples tend to reflect the distribution of scores in the population. If we were interested in using the sample to estimate the population average or to say something about the variability in SAT scores, even the smallest of the samples ($n = 50$) pictured would provide reliable information. Although it is possible to obtain a simple random sample that does not do a reasonable job of representing the population, this is likely only when the sample size is very small, and unless the population itself is small, this risk does not depend on what fraction of the population is sampled. The random selection process inherent in producing a simple random sample allows us to be confident that the sample adequately reflects the population, even when the sample consists of only a small fraction of the population.



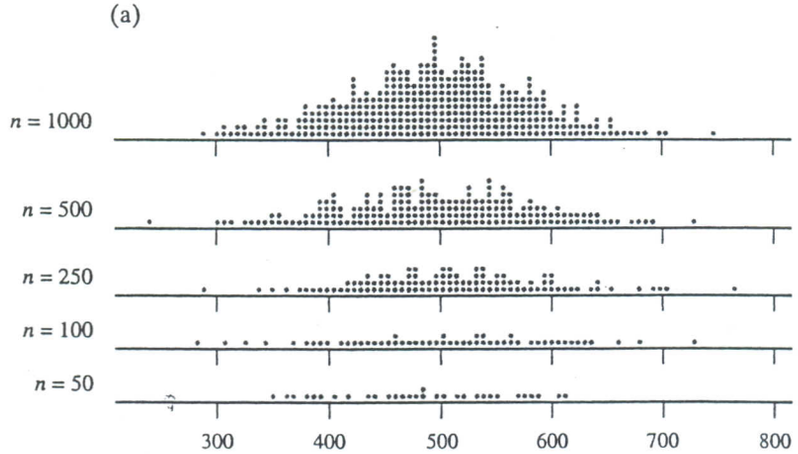


Figure 2.1 (a) Dotplot of math SAT scores for the entire population. (b) Dotplots of math SAT scores for random samples of sizes 50, 100, 250, 500, and 1000.

▪ Other Sampling Methods

Simple random sampling provides researchers with a sampling method that is objective and free of selection bias. In some settings, however, alternative sampling methods may be less costly, easier to implement, or more accurate.

▪ **Stratified Random Sampling** When the entire population can be divided into a set of nonoverlapping subgroups, a method known as **stratified sampling** often proves easier to implement and more cost-effective than random sampling. In stratified random sampling, separate simple random samples are independently selected from each subgroup. For example, to estimate the average cost of malpractice insurance, a researcher might find it convenient to view the population of all doctors practicing in a particular metropolitan area as being made up of four subpopulations: (1) surgeons, (2) internists and family practitioners, (3) obstetricians, and (4) a group that includes all other areas of specialization. Rather than taking a simple random sample from the population of all doctors, the researcher could take four separate random samples — one from the group of surgeons, another from the internists and family practitioners, and so on. These four samples would provide information about the four subgroups as well as information about the overall population of doctors.

When the population is divided in this way, the subgroups are called **strata**. Stratified sampling entails selecting a separate simple random sample from each stratum. Stratified sampling can be used instead of random sampling if it is important to obtain information about characteristics of the individual strata as well as of the entire population, although a stratified sample is not required to do this — subgroup estimates can also be obtained by using an appropriate subset of data from a simple random sample.

The real advantage of stratified sampling is that it often allows us to make more accurate inferences about a population than does simple random sampling. In general, it is much easier to estimate characteristics of a homogeneous group than of a heterogeneous group. For example, even with a small sample, it is possible to obtain an accurate estimate of the average grade point average (GPA) of students graduating with high honors from a university. These students are a homogeneous group with respect to GPA. The individual GPAs of group members are all quite similar, and even a sample of three or four individuals from this subpopulation should be representative. On the other hand, estimating the average GPA of *all* seniors at the university, a much more diverse group of GPAs, is a more difficult task.

If a varied population can be divided into strata, with each stratum being much more homogeneous than the population with respect to the characteristic of interest, then stratified sampling tends to produce more accurate estimates of population characteristics than simple random sampling does. This is because when the strata are relatively homogeneous, a small sample from each stratum can provide reasonably accurate information about strata characteristics. The strata information can then be used to obtain better information about the population as a whole than might have been possible using a simple random sample of the same total size.

▪ **Cluster Sampling** Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves. For example, suppose that a large urban high school has 600 senior students, all of whom are enrolled in a first period homeroom. There are 24 senior homerooms, each with approximately 25 students. If school administrators wanted to select a sample of roughly 75 seniors to participate in an evaluation of the college and career placement advising available to students, they might find it much easier to select three of the senior homerooms at random and then include all the students in the selected homerooms in the sample. In this way, an evaluation survey could be administered to all students in the selected homerooms at the same time — certainly easier logistically than randomly selecting students and then administering the survey to the 75 individual seniors.

Cluster sampling involves dividing the population of interest into nonoverlapping subgroups, called **clusters**. Clusters are then selected at random, and all individuals in the selected clusters are included in the sample.

Because whole clusters are selected, the ideal situation occurs when each cluster mirrors the characteristics of the population. When this is the case, a small number of clusters results in a sample that is representative of the population. If it is not reasonable to think that the variability present in the population is reflected in each cluster, as is often the case when the cluster sizes are small, then it becomes important to ensure that a large number of clusters is included in the sample.

Be careful not to confuse clustering and stratification. Even though both of these sampling strategies involve dividing the population into subgroups, both the way in which the subgroups are sampled and the optimal strategy for creating the subgroups are different. In stratified sampling, we sample from every stratum, whereas in cluster sampling, only selected whole clusters are included in the sample. Because of this difference, to increase the chance of obtaining a sample that is representative of the population, we want to create homogeneous (similar) groups for strata and heterogeneous (reflecting the variability in the population) groups for clusters.

▪ **Systematic Sampling** **Systematic sampling** is a procedure that can be used when it is possible to view the population of interest as consisting of a list or some other sequential arrangement. A value k is specified (e.g., $k = 50$ or $k = 200$). Then one of the first k individuals is selected at random, after which every kth individual in the sequence is included in the sample. A sample selected in this way is called a **1 in k systematic sample**.

For example, a sample of faculty members at a university might be selected from the faculty phone directory. One of the first 20 faculty members listed could be selected at random, and then every 20th faculty member after that on the list would also be included in the sample. This would result in a 1 in 20 systematic sample.

The value of k for a 1 in k systematic sample is generally chosen to achieve a desired sample size. For example, in the faculty directory scenario just described, if there were 900 faculty members at the university, the 1 in 20 systematic sample described would result in a sample size of 45. If a sample size of 100 was desired, a 1 in 9 systematic sample could be used.

As long as there are no repeating patterns in the population list, systematic sampling works reasonably well. The potential danger is that if there are such patterns, systematic sampling can result in an unrepresentative sample.

▪ **Convenience Sampling: Don't Go There!** It is often tempting to resort to "convenience" sampling — that is, using an easily available or convenient group to form a sample. This is a recipe for disaster! Results from such samples are rarely informative, and it is a mistake to try to generalize from a convenience sample to any larger population.

One common form of convenience sampling is sometimes called voluntary response sampling. Such samples rely entirely on individuals who volunteer to be a part of the sample, often by responding to an advertisement, calling a publicized telephone number to register an opinion, or logging on to an Internet site to complete a survey. It is extremely unlikely that individuals participating in such voluntary response surveys are representative of any larger population of interest.

▪ Exercises 2.1–2.26

2.1 The psychology department at a university graduated 140 students in 2003. As part of a curriculum review, the department would like to select a simple random sample of twenty 2003 graduates to obtain information on how graduates perceived the value of the curriculum. Describe two different methods that might be used to select the sample.

2.2 During the previous calendar year, a county's small claims court processed 870 cases. A legal researcher would like to select a simple random sample of 50 cases to obtain information regarding the average award in such cases. Describe how a simple random sample of size $n = 50$ might be selected from the case files.

2.3 A petition with 500 signatures is submitted to a university's student council. The council president would like to determine the proportion of those who signed the petition who are actually registered students at the university. There is not enough time to check all 500 names with the registrar, so the council president decides to select a simple random sample of 30 signatures. Describe how this might be done.

2.4 The financial aid officers of a university wish to estimate the average amount of money that students spend on textbooks each term. They are considering taking a stratified sample. For each of the following proposed stratification schemes, discuss whether you think it would be worthwhile to stratify the university students in this manner.

a. Strata corresponding to class standing (freshman, sophomore, junior, senior, graduate student)

b. Strata corresponding to field of study, using the following categories: engineering, architecture, business, other

c. Strata corresponding to the first letter of the last name: A–E, F–K, etc.

2.5 Citrus trees are usually grown in orderly arrangements of rows to facilitate automated farming and harvesting practices. Suppose that a group of 1000 trees is laid out in 40 rows of 25 trees each. To determine the sugar content of fruit from a sample of 30 trees, researcher A suggests randomly selecting five rows and then randomly selecting six trees from each sampled row. Researcher B suggests numbering each tree on a map of the trees from 1 to 1000 and using random numbers to select 30 of the trees. Is there a reason for preferring one of these sample selection methods to the other? Explain.

2.6 For each of the situations described in what follows, state whether the sampling procedure is simple random sampling, stratified random sampling, cluster sampling, systematic sampling, or convenience sampling.

a. All freshmen at a university are enrolled in 1 of 30 sections of a freshman seminar course. To select a sample of freshmen at this university, a researcher selects 4 sections of the freshman seminar course at random from the 30 sections and all students in the 4 selected sections are included in the sample.

b. To obtain a sample of students, faculty, and staff at a university, a researcher randomly selects 50 faculty members from a list of faculty, 100 students from a list of students, and 30 staff members from a list of staff.

c. A university researcher obtains a sample of students at his university by using the 85 students enrolled in his Psychology 101 class.

d. To obtain a sample of the seniors at a particular high school, a researcher writes the name of each senior on a slip of paper, places the slips in a box and mixes them, and then selects 10 slips. The students whose names are on the selected slips of paper are included in the sample.

e. To obtain a sample of those attending a basketball game, a researcher selects the 24th person through the door. Then, every 50th person after that is also included in the sample.

2.7 A small private college has 4500 students enrolled. Assume that the university can provide a list of the students with the students numbered from 1 to 4500. Describe the procedure you will use to select a simple random sample of 20 students, and then identify (by number) which students from the list are included in your sample.

2.8 Of the 6500 students enrolled at a community college, 3000 are part-time and the other 3500 are enrolled full time. Assume that the college can provide you with a list of students that is sorted so that all full-time students are listed first, followed by the part-time students.

a. Select a stratified random sample that uses full-time and part-time students as the two strata and that includes 10 students from each stratum. Describe the procedure you used to select the sample, and identify the students included in your sample by placement on the sorted list.

b. Does every student at this community college have the same chance of being selected for inclusion in the sample? Explain.

2.9 Give a brief explanation of why it is advisable to avoid the use of convenience samples.

2.10 Sometimes samples are composed entirely of volunteer responders. Give a brief description of the dangers of using voluntary response samples.

2.11 For no apparent reason, the authors of this textbook would like to know something about the number of words on individual pages of this book. A sample of pages from the book is to be obtained, and the number of words on each selected page will be determined. For the purposes of this exercise, equations are not counted as words and a number is counted as a word only if it is spelled out – that is, *ten* is counted as a word, but *10* is not.

a. Describe a sampling procedure that would result in a simple random sample of pages from this book.
b. Describe a sampling procedure that would result in a stratified random sample. Explain why you chose the specific strata used in your sampling plan.
c. Describe a sampling procedure that would result in a systematic sample.
d. Describe a sampling procedure that would result in a cluster sample.
e. Using the process you gave in Part (a), select a simple random sample of at least 20 pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.
f. Using the process you gave in Part (b), select a stratified random sample that includes a total of at least 20 selected pages, and record the number of words on each of the selected pages. Construct a dotplot of the resulting sample values, and write a sentence or two commenting on what it reveals about the number of words on a page.

2.12 In 2000, the chairman of a California ballot initiative campaign to add "none of the above" to the list of ballot options in all candidate races was quite critical of a Field poll that showed his measure trailing by 10 percentage points. The poll was based on a random sample of 1000 registered voters in California. He is quoted by the Associated Press (January 30, 2000) as saying, "Field's sample in that poll equates to one out of 17,505 voters," and he added that this was so dishonest that Field should get out of the polling business! If you worked on the Field poll, how would you respond to this criticism?

2.13 A pollster for the Public Policy Institute of California explains how the Institute selects a sample of California adults ("It's About Quality, Not Quantity," *San Luis Obispo Tribune*, January 21, 2000):

> That is done by using computer-generated random residential telephone numbers with all California prefixes, and when there are no answers, calling back repeatedly to the original numbers selected to avoid a bias against hard-to-reach people. Once a call is completed, a second random selection is made by asking for the adult in the household who had the most recent birthday. It is as important to randomize who you speak to in the household as it is to randomize the household you select. If you didn't, you'd primarily get women and older people.

Comment on this approach to selecting a sample. How does the sampling procedure attempt to minimize certain types of bias? Are there sources of bias that may still be a concern?

2.14 "More than half of California's doctors say they are so frustrated with managed care they will quit, retire early, or leave the state within three years." This statement is the lead sentence in an article titled "Doctors Feeling Pessimistic, Study Finds" (*San Luis Obispo Tribune*, July 15, 2001). This conclusion was based on a mail survey conducted by the California Medical Association. Surveys were mailed to 19,000 California doctors, and 2000 completed surveys were returned. Describe any concerns you have regarding the conclusion drawn in the first sentence of the article.

2.15 The article "I'd Like to Buy a Vowel, Drivers Say" (*USA Today*, August 7, 2001) speculates that young people prefer automobile names that consist of just numbers and/or letters that do not form a word (such as Hyundai's XG300, Mazda's 626, and BMW's 325i). The article goes on to state that Hyundai had planned to identify the car now marketed as the XG300 with the name Concerto, until they determined that consumers hated it and that they thought XG300 sounded more "technical" and deserving of a higher price. Do the students at your school feel the same way? Describe how you would go about selecting a sample to answer this question.

2.16 Based on a survey of 4113 U.S. adults, researchers at Stanford University concluded that Internet use leads to increased social isolation. The survey was conducted by an Internet-based polling company that selected its samples from a pool of 35,000 potential respondents, all of whom had been given free Internet access and WebTV hardware in exchange for agreeing to regularly participate in surveys conducted by the polling company. Two criticisms of this study were expressed in an article that appeared in the *San Luis Obispo Tribune* (February 28, 2000). The first criticism was that increased social isolation was measured by asking respondents if they were talking less to family and friends on the phone. The second criticism was that the sample was

selected only from a group that was induced to participate by the offer of free Internet service, yet the results were generalized to all U.S. adults. Each of these criticisms is describing a type of bias. For each one, indicate what type of bias is being described and why it might make you question the conclusion drawn by the researchers.

2.17 "Workers Grow More Dissatisfied" is the headline for an article that appeared in the *San Luis Obispo Tribune* (August 22, 2002). The article states that "a survey of 5000 people found that while most Americans continue to find their jobs interesting, and are even satisfied with their commutes, a bare majority like their jobs." This statement was based on the fact that only 51 percent of those responding to the survey indicated that they were satisfied with their jobs. The survey was conducted by mail. Describe any potential sources of bias that you think might limit the researcher's ability to draw conclusions about working Americans based on the data collected in this survey.

2.18 *USA Today* (October 4, 2002) gave the accompanying summary data for private golf clubs in nine geographic regions of the United States in an article titled "If Women Want to Start Their Own Golf Clubs, More Power to Them." The value reported for each region is the median annual dues for the private clubs in the region that responded to the survey. The median is a middle value in the sense that half the clubs responding had dues that were lower and half had dues that were higher than the reported median. A total of 2900 private golf clubs were surveyed, with 11% responding:

Region	1	2	3	4	5
Median Dues	5200	1860	3680	2580	2950
Region	6	7	8	9	
Median Dues	2680	2400	9800	2040	

The article states that "private clubs in the U.S. are just that: Most decline to reveal their membership criteria or fees. But a general sense of how much their estimated 6 million members pay can be gleaned from this survey of private clubs." Do you agree that it is reasonable to use the reported values to get a sense of what the annual dues are in the nine geographic regions? Explain why or why not.

2.19 Studies linking abnormal genes to an increased risk of breast cancer are now being reevaluated. The article "Gene's Role in Cancer May Be Overstated" (*San Luis Obispo Tribune*, August 21, 2002) states that "early studies that evaluated breast cancer risk among gene mutation carriers selected women in families where sisters, mothers, and grandmothers all had breast cancer. This created a statistical bias that skewed risk estimates for women in the general population." Is the bias described here selection bias, measurement bias, or nonresponse bias? Explain.

2.20 If liquor stores accept credit cards, does alcohol consumption change? The article "Changes in Alcohol Consumption Patterns Following the Introduction of Credit Cards in Ontario Liquor Stores" (*Journal of Studies on Alcohol* [1999]: 378–382) attempted to answer this question with data based on a telephone survey of adults. The sample was obtained by generating random phone numbers.

a. What are some potential drawbacks of using random phone numbers for sampling? Do any of these drawbacks seem critical for answering this question about alcohol and credit cards?

b. What time of day would be best for such calls?

2.21 The article "Drinking-Driving and Riding with Drunk Drivers Among Young Adults: An Analysis of Reciprocal Effects" (*Journal of Studies on Alcohol* [1999]: 615–621) concludes that although 16–24-year-olds who drive drunk are more likely to ride with drunk drivers, those who ride with drunk drivers are not more likely to drive drunk themselves. These results are based on a sample of 993 young adults from the state of New York, excluding New York City residents. The article explains the exclusion of New York City by saying that the proportion of young adults is higher but that the driving rate of young adults is far lower.

a. Describe the population of interest for this study.

b. Explain how this sample would (or would not) be representative of New York City young adults.

c. Explain how this sample would (or would not) be representative of adults from the entire United States.

2.22 The study described in Exercise 2.21 actually used a stratified sample with strata determined by age. Why do you think the stratification was desirable?

2.23 According to the article "Effect of Preparation Methods on Total Fat Content, Moisture Content, and Sensory Characteristics of Breaded Chicken Nuggets and Beef Steak Fingers" (*Family and Consumer Sciences Research Journal* [1999]: 18–27), sensory tests were conducted using 40 college student volunteers at Texas Women's University. Give three reasons, apart from the relatively small sample size, why this sample may not be ideal as the basis for generalizing to the population of all college students.

2.24 According to an article on the CNN.com web site (dated September 17, 1998) titled "Majority of U.S. Teens Are Not Sexually Active, Study Shows," 52% of surveyed teenagers had never had sexual in-

tercourse. A very large random sample of 16,262 high school students was the source of this information. If the population of interest consists of all teenagers in the United States, are there individuals in the population who had no chance of being included in the sample? What is the name for this type of bias?

2.25 A new and somewhat controversial polling procedure that replaces the phone with the Internet is being used to conduct public opinion polls. The article "Pollsters Fear Internet Numbers Offer Warped View" (*Knight Ridder Newspapers*, October 18, 1999) states that "many pollsters say Internet users as a whole are still too upper-income, highly educated, white, and urban to produce results that accurately reflect all Americans. What's more, while 94% of U.S. households had a telephone in 1998, only 26% had Internet access, according to a U.S. Department of Commerce study." What type of bias is being described in this statement? Do you think that this bias is a serious problem? Explain.

2.26 The article "Study Provides New Data on the Extent of Gambling by College Athletes" (*Chronicle of Higher Education*, January 22, 1999) reported that "72 percent of college football and basketball players had bet money at least once since entering college." This conclusion was based on a study in which "copies of the survey were mailed to 3000 athletes at 182 Division I institutions, 25 percent of whom responded." What types of bias might have influenced the results of this study? Explain.

▪ 2.3 Statistical Studies: Observation and Experimentation

During the week of August 10, 1998, articles with the following headlines appeared in *USA Today*:

"Prayer Can Lower Blood Pressure" (August 11, 1998)

"Wired Homes Watch 15% Less Television" (August 13, 1998)

In each of these articles, a conclusion is drawn from data. In the study on prayer and blood pressure, 2391 people, 65 years or older, were followed for 6 years. The article states that people who attended a religious service once a week and prayed or studied the Bible at least once a day were less likely to have high blood pressure. The researcher then concluded that "attending religious services lowers blood pressure."

The second article reported on a study of 5000 families, some of whom had online Internet services. The researcher concluded that TV use was lower in homes that were wired for the Internet, compared to nonwired homes. The researcher who conducted the study cautioned that the study does not answer the question of whether higher family income could be another factor in lower TV usage.

Are these conclusions reasonable? As we will see, the answer in each case depends in a critical way on how the data were collected.

▪ Observation and Experimentation

Data collection is an important step in the data analysis process. When we set out to collect information, it is important to keep in mind the questions we hope to answer on the basis of the resulting data. Sometimes we are interested in answering questions about characteristics of an existing population or in comparing two or more well-defined populations. To accomplish this, a sample is selected from each population under consideration, and the sample information is used to gain insight into characteristics of the population(s).

For example, an ecologist might be interested in estimating the average shell thickness of bald eagle eggs. A social scientist studying a rural community may want to determine whether gender and attitude toward abortion are related. A city

council member in a college town may want to ascertain whether student residents of the city differ from nonstudent residents with respect to their support for various community projects. These are examples of studies that are *observational* in nature. We want to observe characteristics of members of an existing population or of several populations, and then use the resulting information to draw conclusions. In an **observational study**, it is important to obtain a sample that is representative of the corresponding population. To be reasonably certain of this, the researcher must carefully consider the way in which the sample is selected.

Sometimes the questions we are trying to answer deal with the effect of certain explanatory variables on some response and cannot be answered using data from an observational study. Such questions are often of the form, What happens when . . . ? or, What is the effect of . . . ? For example, an educator may wonder what would happen to test scores if the required lab time for a chemistry course were increased from 3 hr to 6 hr per week. To answer questions such as this one, the researcher conducts an experiment to collect relevant data. The value of some response variable (test score in the chemistry example) is recorded under different experimental conditions (3-hr lab and 6-hr lab in the chemistry example). In an experiment, the researcher manipulates one or more variables, called **factors**, to create the experimental conditions.

A study is an **observational study** if the investigator observes characteristics of a subset of the members of one or more existing populations. The goal of an observational study is usually to draw conclusions about the corresponding population or about differences between two or more populations.

A study is an **experiment** if the investigator observes how a response variable behaves when the researcher manipulates one or more factors. The usual goal of an experiment is to determine the effect of the manipulated factors on the response variable. In a well-designed experiment, the researcher determines the composition of the groups that will be exposed to different experimental conditions by random assignment to groups.

The type of conclusion that can be drawn from a research study depends on the study design. Both observational studies and experiments can be used to compare groups, but in an experiment the researcher controls who is in which group, whereas this is not the case in an observational study. This seemingly small difference is critical when it comes to the interpretation of results from the study.

A well-designed experiment can result in data that provide evidence for a cause-and-effect relationship. This is an important difference between an observational study and an experiment. In an observational study, it is impossible to draw cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the factor being studied. Such variables are called **confounding variables**.

A **confounding variable** is one that is related to both group membership and the response variable of interest in a research study.

Both of the studies reported in the *USA Today* articles described at the beginning of this section were observational studies. In the TV viewing example, the researcher merely observed the TV viewing behavior of individuals in two groups

(wired and nonwired) but did not control which individuals were in which group. A possible confounding variable here is family income. Although it may be reasonable to conclude that the wired and nonwired groups differ with respect to TV viewing, it is not reasonable to conclude that the lower TV use in wired homes is caused by the fact that they have Internet service. In the prayer example, two groups were compared (those who attend a religious service at least once a week and those who do not), but the researcher did not manipulate this factor to observe the effect on blood pressure by assigning people at random to service-attending or non-service-attending groups. As a result, cause-and-effect conclusions such as "prayer can lower blood pressure" are not reasonable based on the observed data. It is possible that some other variable, such as lifestyle, is related to both religious service attendance and blood pressure. In this case, lifestyle is an example of a potential confounding variable.

▪ Drawing Conclusions from Statistical Studies

In this section, two different types of conclusions have been described. One type involves generalizing from what we have seen in a sample to some larger population, and the other involves reaching a cause-and-effect conclusion about the effect of an explanatory variable on a response. When is it reasonable to draw such conclusions? The answer depends on the way that the data were collected. Table 2.1 summarizes the types of conclusions that can be made with different study designs.

Table 2.1 ▪ Drawing Conclusions from Statistical Studies

Study Description	Reasonable to Generalize Conclusions about Group Characteristics to the Population?	Reasonable to Draw Cause-and-Effect Conclusion?
Observational study with sample selected at random from population of interest	Yes	No
Observational study based on convenience or voluntary response sample (poorly designed sampling plan)	No	No
Experiment with groups formed by random assignment of individuals or objects to experimental conditions		
Individuals or objects used in study are volunteers or not randomly selected from some population of interest	No	Yes NO
Individuals or objects used in study are randomly selected from some population of interest	Yes	Yes
Experiment with groups not formed by random assignment to experimental conditions (poorly designed experiment)	No	No

As you can see from Table 2.1, it is important to think carefully about the objectives of a statistical study before planning how the data will be collected. Neither type of conclusion (generalizing to a population or cause-and-effect) can be made from an observational study with a poorly designed sampling plan or from an experiment that does not create experimental groups by random assignment. Both observational studies and experiments must be carefully designed if the resulting data are to be useful. We have already considered the common sampling procedures used in observational studies. In Sections 2.4 and 2.5, we consider experimentation and explore what constitutes good practice in the design of simple experiments.

▪ Exercises 2.27–2.37

2.27 The Associated Press (December 10, 1999) reported on an investigation that concluded that women who suffer severe morning sickness early in pregnancy are more likely to have a girl. This conclusion was reached by researchers in Sweden based on a "scientific study." How do you think the researchers might have collected data that would have enabled them to reach such a conclusion? Do you think that the scientific study referred to in the article was an experiment or an observational study? Explain.

2.28 An article titled "Guard Your Kids Against Allergies: Get Them a Pet" (*San Luis Obispo Tribune*, August 28, 2002) described a study that led researchers to conclude that "babies raised with two or more animals were about half as likely to have allergies by the time they turned six."

a. Do you think this study was an observational study or an experiment? Explain.

b. Describe a potential confounding variable that illustrates why it is unreasonable to conclude that being raised with two or more animals is the cause of the observed lower allergy rate.

2.29 Researchers at the Hospital for Sick Children in Toronto compared babies born to mothers with diabetes to babies born to mothers without diabetes ("Conditioning and Hyperanalgesia in Newborns Exposed to Repeated Heel Lances," *Journal of the American Medical Association* [2002]: 857–861). Babies born to mothers with diabetes have their heels pricked numerous times during the first 36 hours of life in order to obtain blood samples to monitor blood sugar level. The researchers noted that the babies born to diabetic mothers were more likely to grimace or cry when having blood drawn than the babies born to mothers without diabetes. This led the researchers to conclude that babies who experience pain early in life become highly sensitive to pain. Comment on the appropriateness of this conclusion.

2.30 Based on a survey conducted on the DietSmart.com web site, investigators concluded that women who regularly watched *Oprah* were only one-seventh as likely to crave fattening foods as those who watched other daytime talk shows (*San Luis Obispo Tribune*, October 14, 2000).

a. Is it reasonable to conclude that watching *Oprah* causes a decrease in cravings for fattening foods? Explain.

b. Is it reasonable to generalize the results of this survey to all women in the United States? To all women who watch daytime talk shows? Explain why or why not.

2.31 "Crime Finds the Never Married" is the conclusion drawn in an article from *USA Today* (June 29, 2001). This conclusion is based on data from the Justice Department's National Crime Victimization Survey, which estimated the number of violent crimes per 1000 people, 12 years of age or older, to be 51 for the never married, 42 for the divorced or separated, 13 for married individuals, and 8 for the widowed. Does being single cause an increased risk of violent crime? Describe a potential confounding variable that illustrates why it is unreasonable to conclude that a change in marital status causes a change in crime risk.

2.32 The paper "Prospective Randomized Trial of Low Saturated Fat, Low Cholesterol Diet During the First Three Years of Life" (*Circulation* [1996]: 1386–1393) describes an experiment to examine the effect of modifying fat intake in early childhood on blood cholesterol level. The article states that in this experiment "1062 infants were randomized to either the intervention or control group at 7 months of age. The families of the 540 intervention group children were counseled to reduce the child's intake of saturated fat and cholesterol but to ensure adequate energy intake. The control children consumed an unrestricted diet."

The researchers concluded that the blood cholesol level was lower for children in the intervention up. Is it reasonable to conclude that the parental nseling and subsequent reduction in dietary fat l cholesterol are the cause of the reduction in od cholesterol level? Explain why or why not. Do you think it is reasonable to generalize the ults of this experiment to all children? Explain.

3 *USA Today* (January 29, 2003) summarized ta from a survey of Americans with household comes of $75,000 or more. It was reported that % of affluent Americans would rather have more ne than more money.

What condition on how the data were collected uld make the generalization from the sample to e population of affluent Americans reasonable?

Would it be reasonable to generalize from the mple to say that 57% of all Americans would rather ve more time than more money? Explain.

34 "Attending Church Found Factor in Longer ife" is the title of an article that appeared in *USA oday* (August 9, 1999). Based on a "nationally epresentative survey of 3617 Americans," the arcle concludes that "attending services extends ne life span about as much as moderate exercise r not smoking." Comment on the validity of this onclusion.

.35 An article that appeared in the *San Luis Obispo Tribune* (November 11, 1999) was titled "Study oints Out Dangerous Side to SUV Popularity: Half f All 1996 Ejection Deaths Occur in SUVs." This rticle states that sports utility vehicles (SUVs) have much higher rate of passengers being thrown from window during an accident than do automobiles. The article also states that more than half of all leaths caused by ejection involved SUVs — the basis for the conclusion that SUVs are more dangerous than cars. Later in the article, there is a comment that about 98% of those injured or killed in ejection accidents were not wearing seat belts. Comment on the conclusion that SUVs are more dangerous than cars.

2.36 Does living in the South cause high blood pressure? Data from a group of 6278 whites and blacks questioned in the Third National Health and Nutritional Examination Survey between 1988 and 1994 (see CNN.com web site article of January 6, 2000, titled "High Blood Pressure Greater Risk in U.S. South, Study Says") indicates that a greater percentage of Southerners have high blood pressure than do people in any other region of the United States. This difference in rate of high blood pressure was found in every ethnic group, gender, and age category studied. List at least two possible reasons we cannot conclude that living in the South causes high blood pressure.

2.37 Does eating broccoli reduce the risk of prostate cancer? According to an observational study from the Fred Hutchinson Cancer Research Center (see CNN.com web site article titled "Broccoli, Not Pizza Sauce, Cuts Cancer Risk, Study Finds," January 5, 2000), men who ate more cruciferous vegetables (broccoli, cauliflower, brussels sprouts, and cabbage) had a lower risk of prostate cancer. This study made separate comparisons for men who ate different levels of vegetables. According to one of the investigators, "at any given level of total vegetable consumption, as the percent of cruciferous vegetables increased, the prostate cancer risk decreased." Based on this study, is it reasonable to conclude that eating cruciferous vegetables causes a reduction in prostate cancer risk? Explain.

▪ 2.4 Simple Comparative Experiments

Sometimes the questions we are trying to answer deal with the effect of certain explanatory variables on some response. Such questions are often of the form, What happens when . . . ? or, What is the effect of . . . ? For example, an educator may wonder what happens to test scores if an activity-based method of instruction is used rather than a traditional lecture method. An industrial engineer may be considering two different workstation designs and might want to know whether the choice of design affects work performance. A medical researcher may want to determine how a proposed treatment for a disease compares to a standard treatment. A nutritionist working for a commercial bakery may want to determine the effect of baking time and baking temperature on the nutritional content of bread.

To address these types of questions, the researcher conducts an experiment to collect the relevant information. The value of some response variable (test score, assembly time, bread density, etc.) is determined under different experimental conditions. Experiments must be carefully planned to obtain information that will give unambiguous answers to questions of interest.

▪ Definition

An **experiment** is a planned intervention undertaken to observe the effects of one or more explanatory variables, often called **factors**, on a response variable. The fundamental purpose of the intervention is to increase understanding of the nature of the relationships between the explanatory and response variables. Any particular combination of values for the explanatory variables is called an **experimental condition** or **treatment**.

The **design** of an experiment is the overall plan for conducting an experiment. A good design minimizes ambiguity in the interpretation of the results.

Suppose that we are interested in determining how student performance on a first-semester calculus exam is affected by room temperature. There are four sections of calculus being offered in the fall semester. We might design an experiment in this way: Set the room temperature (in degrees Fahrenheit) to 65° in two of the rooms and to 75° in the other two rooms on test day, and then compare the exam scores for the 65° group and the 75° group. Suppose that the average exam score for the students in the 65° group was noticeably higher than the average for the 75° group. Could we conclude that the increased temperature resulted in a lower average score? The answer is no, because many other factors might be related to exam score. Were the sections at different times of the day? Did they have the same instructor? Different textbooks? Was one section required to do more homework than the other sections? Did the sections differ with respect to the abilities of the students? Any of these other factors could provide a plausible explanation (having nothing to do with room temperature) for why the average test score was different for the two groups. It is not possible to separate the effect of temperature from the effects of these other factors. As a consequence, simply setting the room temperatures as described makes for a poorly designed experiment. A well-designed experiment requires more than just manipulating the explanatory variables; the design must also eliminate rival explanations or else the experimental results will not be conclusive.

The goal is to design an experiment that will allow us to determine the effects of the relevant factors on the chosen response variable. To do this, we must take into consideration other extraneous factors that, although not of interest in the current study, might also affect the response variable.

▪ Definition

An **extraneous factor** is one that is not of interest in the current study but is thought to affect the response variable.

A researcher can **directly control** some extraneous factors. In the calculus test example, the textbook used is an extraneous factor because part of the differences in test results might be attributed to this factor. We might choose to control this factor directly, by requiring that all sections participating in the study use the same textbook. If this were the case, any observed differences between temperature groups could not be explained by the use of different textbooks, because all sections would have used the same book. The extraneous factor *time of day* might also be directly controlled in this way by having all sections meet at the same time.

The effects of some extraneous factors can be filtered out by a process known as **blocking**. Extraneous factors that are addressed through blocking are called *blocking factors*. Blocking creates groups (called blocks) that are similar with respect to blocking factors; then all treatments are tried in each block. In our example, we might use *instructor* as a blocking factor. If two instructors are each teaching two sections of calculus, we would make sure that for each instructor, one section was part of the 65° group and the other section was part of the 75° group. With this design, if we see a difference in exam scores for the two temperature groups, the factor *instructor* can be ruled out as a possible explanation, because both instructors' students were present in each temperature group. If one instructor taught both 65° sections and the other taught both 75° sections, we would be unable to distinguish the effect of temperature from the effect of the instructor. These two factors (temperature and instructor) are said to be **confounded**.

Two factors are **confounded** if their effects on the response variable cannot be distinguished from one another.

We can directly control some extraneous factors by holding them constant, and we can use blocking to create groups that are similar to essentially filter out the effect of others. What about factors, such as student ability in our calculus test example, that cannot be controlled by the experimenter and that would be difficult to use as blocking factors? These extraneous factors are handled by the use of random assignment to experimental groups — a process called **randomization**. Randomization ensures that our experiment does not favor one experimental condition over any other and attempts to create "equivalent" experimental groups (groups that are as much alike as possible). For example, if the students requesting calculus could be assigned to one of the four available sections using some sort of random mechanism, we would expect the resulting groups to be similar with respect to student ability as well as with respect to other extraneous factors that we are not directly controlling or using as a basis for blocking.

To get a sense of how random assignment tends to create similar groups, suppose that 50 first-year college students are available to participate as subjects in an experiment to investigate whether completing an online review of course material before an exam improves exam performance. The 50 subjects vary quite a bit with respect to achievement, which is reflected in their math and verbal SAT scores, as shown in Figure 2.2.

If these 50 students are to be assigned to the two experimental groups (one that will complete the online review and one that will not), we want to make sure that

A researcher can **directly control** some extraneous factors. In the calculus test example, the textbook used is an extraneous factor because part of the differences in test results might be attributed to this factor. We might choose to control this factor directly, by requiring that all sections participating in the study use the same textbook. If this were the case, any observed differences between temperature groups could not be explained by the use of different textbooks, because all sections would have used the same book. The extraneous factor *time of day* might also be directly controlled in this way by having all sections meet at the same time.

The effects of some extraneous factors can be filtered out by a process known as **blocking**. Extraneous factors that are addressed through blocking are called *blocking factors*. Blocking creates groups (called blocks) that are similar with respect to blocking factors; then all treatments are tried in each block. In our example, we might use *instructor* as a blocking factor. If two instructors are each teaching two sections of calculus, we would make sure that for each instructor, one section was part of the 65° group and the other section was part of the 75° group. With this design, if we see a difference in exam scores for the two temperature groups, the factor *instructor* can be ruled out as a possible explanation, because both instructors' students were present in each temperature group. If one instructor taught both 65° sections and the other taught both 75° sections, we would be unable to distinguish the effect of temperature from the effect of the instructor. These two factors (temperature and instructor) are said to be **confounded**.

Two factors are **confounded** if their effects on the response variable cannot be distinguished from one another.

We can directly control some extraneous factors by holding them constant, and we can use blocking to create groups that are similar to essentially filter out the effect of others. What about factors, such as student ability in our calculus test example, that cannot be controlled by the experimenter and that would be difficult to use as blocking factors? These extraneous factors are handled by the use of random assignment to experimental groups — a process called **randomization**. Randomization ensures that our experiment does not favor one experimental condition over any other and attempts to create "equivalent" experimental groups (groups that are as much alike as possible). For example, if the students requesting calculus could be assigned to one of the four available sections using some sort of random mechanism, we would expect the resulting groups to be similar with respect to student ability as well as with respect to other extraneous factors that we are not directly controlling or using as a basis for blocking.

To get a sense of how random assignment tends to create similar groups, suppose that 50 first-year college students are available to participate as subjects in an experiment to investigate whether completing an online review of course material before an exam improves exam performance. The 50 subjects vary quite a bit with respect to achievement, which is reflected in their math and verbal SAT scores, as shown in Figure 2.2.

If these 50 students are to be assigned to the two experimental groups (one that will complete the online review and one that will not), we want to make sure that

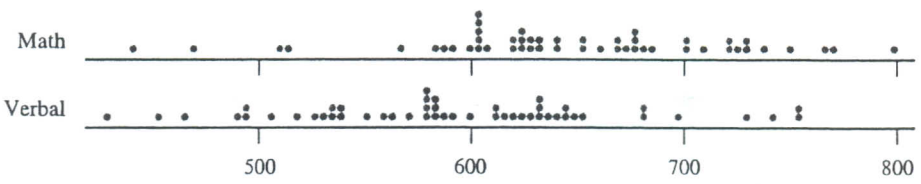


Figure 2.2 Dotplots of math and verbal SAT scores for 50 first-year students.

the assignment of students to groups does not favor one group over the other by tending to assign the higher achieving students to one group and the lower achieving students to the other.

Creating groups of students with similar achievement levels in a way that considers both verbal and math SAT scores simultaneously would be difficult, so we rely on random assignment. Figure 2.3(a) shows the verbal SAT scores of the students assigned to each of the two experimental groups (one shown in orange and one shown in blue) for each of three different random assignments of students to groups. Figure 2.3(b) shows the math SAT scores for the two experimental groups for each of the same three random assignments. Notice that each of the three random assignments produced groups that are similar with respect to *both* verbal and math SAT scores. So, if any of these three assignments were used and the two groups differed on exam performance, we could rule out differences in math or verbal SAT scores as possible competing explanations for the difference.

Not only will random assignment tend to create groups that are similar with respect to verbal and math SAT scores, but it will also tend to even out the groups with respect to other extraneous variables. As long as the number of subjects is not too small, we can rely on the random assignment to produce comparable experimental groups. This is the reason that randomization is a part of all well-designed experiments.

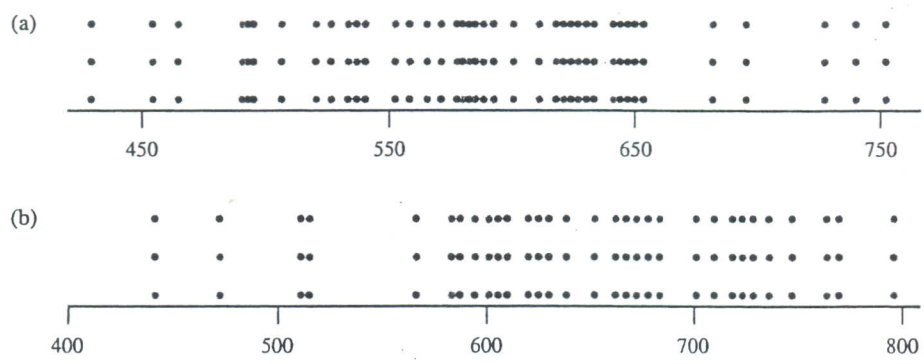


Figure 2.3 Dotplots for three different random assignments to two groups, one shown in orange and one shown in blue: (a) verbal SAT score and (b) math SAT score.

Not all experiments require the use of human subjects to evaluate different experimental conditions. For example, a researcher might be interested in comparing three different gasoline additives with respect to automobile performance, as measured by gasoline mileage. The experiment might involve using a single car with an empty tank. One gallon of gas with one of the additives will be put in the tank, and the car will be driven along a standard route at a constant speed until it runs out of gas. The total distance traveled on the gallon of gas could then be recorded. This could be repeated a number of times — ten, for example — with each additive.

The experiment just described can be viewed as consisting of a sequence of trials. Because a number of extraneous factors might have an effect on gas mileage

(such as variations in environmental conditions — e.g., wind speed or humidity — and small variations in the condition of the car), it would not be a good idea to use additive 1 for the first 10 trials, additive 2 for the next 10 trials, and so on. An approach that would not unintentionally favor any one of the additives would be to randomly assign additive 1 to 10 of the 30 planned trials, and then randomly assign additive 2 to 10 of the remaining 20 trials. The resulting plan for carrying out the experiment might look as follows:

Trial	1	2	3	4	5	6	7	...	30
Additive	2	2	3	3	2	1	2	...	1

When an experiment can be viewed as a sequence of trials, randomization involves the random assignment of treatments to trials. *Just remember that random assignment — either of subjects to treatments or of treatments to trials — is a critical component of a good experiment.*

Randomization can be effective in evening out the effects of extraneous variables only if the number of subjects or observations in each treatment or experimental condition is large enough for each experimental group to reliably reflect variability in the population. For example, if there were only eight students requesting calculus, it is unlikely that we would get equivalent groups for comparison, even with random assignment to the sections. **Replication** is the design strategy of making multiple observations for each experimental condition. Together, replication and randomization allow the researcher to be reasonably confident of comparable experimental groups.

▪ **Key Concepts in Experimental Design**

Randomization
Random assignment (of subjects to treatments or of treatments to trials) to ensure that the experiment does not systematically favor one experimental condition over another.

Blocking
Using extraneous factors to create groups (blocks) that are similar. All experimental conditions are then tried in each block.

Direct Control
Holding extraneous factors constant so that their effects are not confounded with those of the experimental conditions.

Replication
Ensuring that there is an adequate number of observations for each experimental condition.

These four concepts — randomization, blocking, control, and replication — are the fundamental principles of experimental design. When you collect data through experimentation, you should give careful thought to the role that each of these principles plays in your design.

It is an unremarkable event when we read in our newspapers or see on the news or perhaps read in one of our textbooks that a scientist performed an experiment. The subject matter may be a new therapy for arthritis, a promising drug for asthma

control, or a substance that holds promise for longer lasting tires for automobiles, but the stories are similar. A scientist, doctor, or engineer has performed an experiment, and some new or unexpected result has brought new knowledge or refined our understanding of some scientific puzzle or, in some cases, caused rethinking about a heretofore "settled" issue.

To illustrate the design of a simple experiment, consider the dilemma of Anna, a waitress in a local restaurant. Anna is saving money for college, and she depends on tips from her customers in the restaurant. She would like to increase the amount of her tips, and her strategy is simple: In addition to prompt and courteous service, she will project a positive image of herself to her customers by writing "Thank you" on the back of the check before she gives it to the patron. She will write "Thank you" on some of the checks, and on others she will write nothing. She plans to calculate the percentage of the tip as her measure of success (for instance, a 15% tip is common). She will compare the average percentage of the tip calculated from checks with and without the handwritten "Thank you."

Writing "Thank you" consumes a bit more time in her already hectic serving activities, and so she has a stake in the outcome of her experiment. If the extra time does not produce higher tips, she may try a different strategy; in any case it is of value to her to have at least a tentative answer to the question, Will writing "Thank you" produce the desired outcome of higher tips? Anna is untrained in the art of planning experiments, but already she has taken some commonsense steps in the right direction to answer her question.

1. *Identification of a specific problem.* Anna has defined a manageable problem, and collecting the appropriate data is feasible. It should be easy to gather data as a normal part of her work.
2. *Definition of the variables of interest.* Anna wonders whether writing "Thank you" on the customers' bills will have an effect on the amount of her tip. In the language of experimentation, we would refer to the writing of "Thank you" and the not writing of "Thank you" as **treatments** (the two experimental conditions to be compared in the experiment). The two treatments together are the possible values of the **explanatory variable**. The tipping percentage is the **response variable**. The idea behind this terminology is that the outcome of the experiment (the tipping percentage) is a *response* to the treatments *writing "Thank you"* or *not writing "Thank you."* Anna's experiment may be thought of as an attempt to explain the variability in the response variable in terms of its presumed cause, the variability in the explanatory variable. That is, as she manipulates the explanatory variable, she expects the response by her customers to vary in a manner consistent with her theory.
3. *Formulation of the measurement.* The explanatory variable is categorical (writing "Thank you" or not writing "Thank you"), and it can be easily recorded. The response variable, *tipping percentage*, is easily calculated from the tip and the original bill: % tip = (tip/bill total) × 100.
4. *Analysis of the resulting data.* After collecting the data, statistical methods can be used to decide whether the tip sizes tend to be higher for the "Thank you" treatment.

Anna has a good start, but now she must consider the four fundamental design principles.

Replication. Anna cannot run a successful experiment by gathering tipping information on only one person for each treatment. There is no reason to be-

lieve that any single tipping incident is representative of what would happen in other incidents (because customers vary in their tipping practices), and therefore it would be impossible to evaluate the two treatments with only two subjects. To interpret the effects of a particular treatment, she must **replicate** each treatment in the experiment.

Control and Randomization. There are a number of extraneous variables in this example — variables that might have an effect on the size of tip. Some restaurant patrons will be seated near the window with a nice view; some will have to wait for a table, whereas others may be seated immediately; and some may be on a fixed income and cannot afford a large tip. Some of these variables can be directly controlled. For example, Anna may choose to use only window tables in her experiment, thus eliminating table location as a potential confounding variable. Other variables, such as length of wait and customer income, cannot be easily controlled. As a result, it is important that Anna use randomization to decide which of the window tables will be in the "Thank you" group and which will be in the no "Thank you" group. She might do this by flipping a coin as she prepares the check for each window table. If the coin lands with the head side up, she could write "Thank you" on the bill, omitting the "Thank you" when a tail is observed.

Blocking. Suppose that Anna works on both Thursdays and Fridays. Because day of the week might affect tipping behavior, Anna should block on day of the week and make sure that observations for both treatments are made on each of the two days.

▪ Evaluating an Experimental Design

The key concepts of experimental design provide a framework for evaluating an experimental design, as illustrated in Examples 2.4 and 2.5.

▪ Example 2.4 Subliminal Messages

The article "The Most Powerful Manipulative Messages Are Hiding in Plain Sight" (*Chronicle of Higher Education*, January 29, 1999) reported the results of an experiment on priming — the effect of subliminal messages on how we behave. In the experiment, subjects completed a language test in which they were asked to construct a sentence using each word in a list of words. One group of subjects received a list of words related to politeness, and a second group was given a list of words related to rudeness. Subjects were told to complete the language test and then come into the hall and find the researcher so that he could explain the next part of the test. When each subject came into the hall, he or she found the researcher engaged in conversation. The researcher wanted to see whether the subject would interrupt the conversation. The researcher found that 63% of those primed with words related to rudeness interrupted the conversation, whereas only 17% of those primed with words related to politeness interrupted.

If we assume that the researcher randomly assigned the subjects to the two groups, then this study is an experiment that compares two treatments (primed with words related to rudeness and primed with words related to politeness). The response variable, *politeness*, has the values *interrupted conversation* and *did not interrupt conversation*. The experiment uses replication (many subjects in each treat-

ment group) and randomization to control for extraneous variables that might affect the response.

Many experiments compare a group that receives a particular treatment to a **control group** that receives no treatment.

▪ Example 2.5 Health Education and Cholesterol Level

Researchers at the University of North Carolina studied 422 third- and fourth-grade children to determine whether a program of health education and exercise was effective in reducing cholesterol level (*San Luis Obispo Telegram-Tribune*, August 4, 1998). Children with high cholesterol levels were divided into three groups. One group attended classes in healthy nutrition twice a week and exercised three times a week. A second group received individualized instruction from a nurse and exercised three times a week. The third group, a control group, received no specialized instruction and participated in the school's regular physical education classes. After 2 months, cholesterol levels were measured. The researchers concluded that the classroom-based program (Group 1) was most effective in lowering cholesterol. If the researchers had randomly assigned the 422 children to the three experimental conditions (treatments), the study would have been an experiment that used both replication and randomization.

When planning an experiment or evaluating a design developed by someone else, be sure to keep the following basic principles in mind:

1. *Replication.* Design strategy of making multiple observations for each experimental treatment.
2. *Direct control.* Holding the values of extraneous variables constant so that their effect is not potentially confounded with factors in the experiment.
3. *Blocking.* Using extraneous variables to create groups (blocks) that are similar, and then making sure that all treatments are tried in each block.
4. *Randomization. A must!* The strategy for dealing with all extraneous variables not taken into account through direct control or blocking. We count on randomization to create "equivalent" experimental groups.

Before proceeding with an experiment, you should be able to give a satisfactory answer to each of the following 10 questions.

1. What is the research question that data from the experiment will be used to answer?
2. What is the response variable?
3. How will the values of the response variable be determined?
4. What are the factors (explanatory variables) for the experiment?
5. For each factor, how many different values are there, and what are these values?
6. What are the treatments for the experiment?
7. What extraneous variables might influence the response?

8. How does the design incorporate random assignment of subjects to treatments (or treatments to subjects) or random assignment of treatments to trials?
9. For each extraneous variable listed in Question 7, how does the design protect against its potential influence on the response through blocking, direct control, or randomization?
10. Will you be able to answer the research question using the data collected in this experiment?

Exercises 2.38–2.48

.38 The head of the quality control department at printing company is interested in factors that affect ıe strength of the binding on books that the comany produces. In particular, she would like to carry ut an experiment to determine which of three dif-:rent glues results in the greatest binding strength. ılthough they are not of interest in the current in-estigation, other factors thought to affect binding trength are the number of pages in the book and /hether the book is being bound as a paperback or a ardback.

. What is the response variable in this experiment?

. What factor will determine the experimental onditions?

. What two extraneous factors are mentioned in the roblem description? Can you think of other extra-ıeous factors that might be considered?

2.39 Based on observing more than 400 drivers in he Atlanta area, two investigators at Georgia State Jniversity concluded that people exiting parking paces did so more slowly when a driver in another ar was waiting for the space than when no one was vaiting ("Territorial Defense in Parking Lots: Retalation Against Waiting Drivers," *Journal of Applied Social Psychology* [1997]: 821–834). Describe how you might design an experiment to collect informaion that would allow you to determine whether this phenomenon is true for your city. What is the reponse variable for your study? What extraneous factors might have an effect on the response variable, and how does your design control for them?

2.40 In 1993, researchers proclaimed that listening to Mozart could make you smarter. Dubbed the Mozart effect, this conclusion was based on a study that showed college students temporarily gained up to 9 IQ points after listening to a Mozart piano sonata. This research has since been criticized by a number of researchers who have been unable to confirm the result in other similar studies. Suppose that you wanted to see whether there is a Mozart effect for students at your school.

a. Describe how you might design an experiment for this purpose. Sample ?

b. Does your experimental design include direct control of any extraneous variables? Explain.

c. Does your experimental design use blocking? Explain why you did or did not include blocking in your design.

d. What role does randomization play in your design?

2.41 An article from the Associated Press (May 14, 2002) led with the headline "Academic Success Lowers Pregnancy Risk." The article described an evaluation of a program that involved about 350 students at 18 Seattle schools in high crime areas. Some students took part in a program beginning in elementary school in which teachers showed children how to control their impulses, recognize the feelings of others, and get what they want without aggressive behavior. Others did not participate in the program. The study concluded that the program was effective because by the time young women in the program reached age 21, the pregnancy rate among them was 38%, compared to 56% for the women in the experiment who did not take part in the program. Explain why this conclusion is valid only if the women in the experiment were randomly assigned to one of the two experimental groups.

2.42 Do ethnic group and gender influence the type of care that a heart patient receives? The following passage is from the article "Heart Care Reflects Race and Sex, Not Symptoms" (*USA Today*, February 25, 1999, reprinted with permission):

> Previous research suggested blacks and women were less likely than whites and men to get cardiac catheterization or coronary bypass surgery for chest pain or a heart attack. Scientists blamed differences in illness severity, insurance coverage, patient preference, and health care access. The researchers eliminated those differences by videotaping actors — two black men, two black women, two white men, and two white women — describing chest pain from identical scripts. They wore

identical gowns, used identical gestures, and were taped from the same position. Researchers asked 720 primary care doctors at meetings of the American College of Physicians or the American Academy of Family Physicians to watch a tape and recommend care. The doctors thought the study focused on clinical decision-making.

Evaluate this experimental design. Do you think this is a good design or a poor design, and why? If you were designing such a study, what, if anything, would you propose to do differently?

2.43 Four new word-processing software programs are to be compared by measuring the speed with which various standard tasks can be completed. Before conducting the tests, researchers note that the level of a person's computer experience is likely to have a large influence on the test results. Discuss how you would design an experiment that fairly compares the word-processing programs while simultaneously accounting for possible differences in users' computer proficiency.

2.44 A study in Florida is examining whether health literacy classes and using simple medical instructions that include pictures and avoid big words and technical terms can keep Medicaid patients healthier (*San Luis Obispo Tribune*, October 16, 2002). Twenty-seven community health centers are participating in the study. For 2 years, half of the centers will administer standard care. The other centers will have patients attend classes and will provide special health materials that are easy to understand. Explain why it is important for the researchers to assign the 27 centers to the 2 groups (standard care and classes with simple health literature) at random.

2.45 Is status related to a student's understanding of science? The article "From Here to Equity: The Influence of Status on Student Access to and Understanding of Science" (*Culture and Comparative Studies* [1999]: 577–602) described a study on the effect of group discussions on learning biology concepts. An analysis of the relationship between status and "rate of talk" (the number of on-task speech acts per minute) during group work included gender as a blocking variable. Do you think that gender is a useful blocking variable? Explain.

2.46 In an experiment to determine whether suggesting a number of items to purchase (e.g., "limit 12 per person") increases the number of items purchased ("An Anchoring and Adjustment Model of Purchase Quantity Decision," *Journal of Marketing Research* [1999]: 71–81), data were collected at three separate stores. At one store the display of on-sale soup had a sign that said "No limit per person"; at another store the sign read "Limit 4 per person," and at the third store the sign read "Limit 12 per person." On the second and third days of the experiment the signs were switched so that each store had each treatment for one day only.

a. Why did the researchers rotate the three treatments across the three different stores for the three days?

b. The results of this study showed that there were no day-to-day or store-to-store differences in the quantity of soup purchased. Also, in the stores with the 12-item limit, the average number of cans purchased per person was twice that of the no-limit or 4-item limit (7 instead of 3.5). This difference was large enough for the researchers to conclude that the results were not due to chance alone. If a grocery store manager asked you how she should market soup, what would you say?

c. If a grocer asked you how she should market lettuce, would you say (based on the results of this study) that a limit of 12 items per person would increase sales over a no-limit" provision? Explain.

2.47 Does alcohol consumption cause increased cravings for cigarettes? Research at Purdue University suggests this is so (see CNN.com web site article "Researchers Find Link Between Cigarette Cravings and Alcohol," dated June 13, 1997). In an experiment, 60 heavy smokers and moderate drinkers were divided into two groups. One group drank vodka tonics and the other group drank virgin tonics (tonic water alone), but all subjects were told they were drinking vodka tonics. The researchers then measured the level of nicotine cravings (by monitoring heart rate, skin conductance, etc.). Those who had consumed the vodka tonics had 35% more cravings than those who did not. Assuming that the assignment of subjects to the treatment (vodka) and control groups was made at random, do you think there are any confounding factors that would make conclusions based on this experiment questionable?

2.48 The level of lead in children's blood should be kept low, and one way to reduce a child's exposure to lead is to control dust in the home. Researchers at the University of Rochester studied the effects of families' dust control on children with low to mildly elevated blood lead levels ("A Randomized Trial of the Effect of Dust Control on Children's Blood Lead Levels," *Pediatrics* [1996]: 35–40). They designed an experiment in which they randomly assigned families whose children had elevated blood lead levels to two groups: intervention and control. The intervention group was given cleaning supplies and a demonstration of proper cleaning techniques. The control group was given a pamphlet on how to avoid lead exposure.

researchers measured lead levels in the children's od and in the household dust at the beginning of study and also 7 months later.
Identify the population in this experiment.
Discuss the role and importance of randomiza- n in this study.
c. Discuss the purpose of a control group.
d. Identify the treatments and response variable(s).
e. Discuss possible extraneous factors and how they could be controlled.
f. Discuss whether and how blocking should be used in this experiment.

2.5 More on Experimental Design

The previous section covered basic principles for designing simple comparative experiments – control, blocking, randomization, and replication. The goal of an experimental design is to provide a method of data collection that (1) minimizes extraneous sources of variability in the response so that any differences in response for various experimental conditions can be more easily assessed and (2) creates experimental groups that are similar with respect to extraneous variables that cannot be controlled either directly or through blocking.

In this section, we look at some additional considerations that you may need to think about when planning an experiment.

▪ Use of a Control Group

If the purpose of an experiment is to determine whether some treatment has an effect, it is important to include an experimental group that does not receive the treatment. Such a group is called a **control group**. The use of a control group allows the experimenter to assess how the response variable behaves when the treatment is not used. This provides a baseline against which the treatment groups can be compared to determine whether the treatment had an effect.

▪ Example 2.6 Comparing Gasoline Additives

Suppose that an experimenter wants to know whether a gasoline additive increases fuel efficiency (miles per gallon). Such an experiment might use a single car (to eliminate car-to-car variability) and a sequence of trials in which 1 gal of gas is put in an empty tank, the car is driven around a racetrack at a constant speed, and the distance traveled on the gallon of gas is recorded.

To determine whether the additive increases gas mileage, it would be necessary to include a control group of trials where distance traveled was measured when gasoline without the additive was used. The experiment might consist of 24 trials. These trials would then be assigned *at random* to one of the two experimental conditions (additive or no additive).

Even though this experiment consists of a sequence of trials all with the same car, random assignment of trials to experimental conditions is still important because there will always be uncontrolled variability. For example, temperature or other environmental conditions might change over the sequence of trials, the physical condition of the car might change slightly from one trial to another, and so on. Random assignment of experimental conditions to trials will tend to even out the effects of these uncontrollable factors.

Although we usually think of a control group as one that receives no treatment, in experiments designed to compare a new treatment to an existing standard treatment, the term *control group* is sometimes used to describe the group that receives the current standard treatment.

Not all experiments require the use of a control group. For example, many experiments are designed to compare two or more conditions — an experiment to compare density for three different formulations of bar soap or an experiment to determine how oven temperature affects the cooking time of a particular type of cake. However, sometimes you will see a control group included even when the ultimate goal is to compare two or more different treatments. An experiment with two treatments and no control group might allow us to determine whether or not there is a difference between the two treatments and even to assess the magnitude of the difference if one exists, but it would not allow us to assess the individual effect of either treatment. For example, without a control group, we might be able to say that there is no difference in the increase in mileage for two different gasoline additives, but we wouldn't be able to tell if this was because both additives increased gas mileage by a similar amount or because neither additive had any effect on gas mileage.

▪ Use of a Placebo

In experiments that use human subjects, use of a control group may not be enough to determine whether a treatment really does have an effect. People sometimes respond merely to the power of suggestion! For example, consider a study designed to determine whether a particular herbal supplement is effective in promoting weight loss. Suppose that the study is designed with an experimental group that takes the herbal supplement and a control group that takes nothing. It is possible that those who take the herbal supplement and believe that they are taking something that will help them to lose weight may be more motivated and may unconsciously change their eating behavior or activity level, resulting in weight loss.

Although there is debate about the degree to which people respond, many studies have shown that people sometimes respond to treatments with no active ingredients, such as sugar pills or solutions that are nothing more than colored water, and that they often report that such "treatments" relieve pain or reduce symptoms such as nausea or dizziness. So, if an experiment is to enable researchers to determine whether a treatment really has an effect, comparing a treatment group to a control group may not be enough. To address the problem, many experiments use what is called a placebo.

A **placebo** is something that is identical (in appearance, taste, feel, etc.) to the treatment received by the treatment group, except that it contains no active ingredients.

For example, in the herbal supplement example, rather than using a control group that received *no* treatment, the researchers might want to include a placebo group. Individuals in the placebo group would take a pill that looked just like the

herbal supplement but did not contain the herb or any other active ingredient. As long as the subjects did not know whether they were taking the herb or the placebo, the placebo group would provide a better basis for comparison and would allow the researchers to determine whether the herbal supplement had any real effect over and above the "placebo effect."

▪ Single-Blind and Double-Blind Experiments

Because people often have their own personal beliefs about the effectiveness of various treatments, it is desirable to conduct experiments in such a way that subjects do not know what treatment they are receiving. For example, in an experiment comparing four different doses of a medication for relief of headache pain, someone who knows that he is receiving the medication at its highest dose may be subconsciously influenced to report a greater degree of headache pain reduction. By ensuring that subjects are not aware of which treatment they receive, we can prevent personal perception from influencing the response.

An experiment in which subjects do not know what treatment they have received is described as **single-blind**. Of course, not all experiments can be made single-blind. For example, in an experiment to compare the effect of two different types of exercise on blood pressure, it is not possible for participants to be unaware of whether they are in the swimming group or the jogging group! However, when it is possible, "blinding" the subjects in an experiment is generally a good strategy.

In some experiments, someone other than the subject is responsible for measuring the response. To ensure that the person measuring the response does not let personal beliefs influence the way in which the response is recorded, the researchers should make sure that the measurer does not know which treatment was given to any particular individual. For example, in a medical experiment to determine whether a new vaccine reduces the risk of getting the flu, doctors must decide whether or not a particular individual who is not feeling well actually has the flu or some other unrelated illness. If the doctor who had to make this assessment knew that a participant with flulike symptoms had been vaccinated with the new flu vaccine, she might be less likely to determine that the participant had the flu and more likely to interpret the symptoms as being the result of some other illness.

There are two ways in which blinding might occur in an experiment. One involves blinding the participants, and the other involves blinding the individuals who measure the response. If participants do not know which treatment was received and those measuring the response do not know which treatment was given to which participant, the experiment is described as **double-blind**. If only one of the two types of blinding is present, the experiment is single-blind.

A **double-blind** experiment is one in which neither the subjects nor the individuals who measure the response know which treatment was received.

A **single-blind** experiment is one in which the subjects do not know which treatment was received, but the individuals measuring the response do know which treatment was received; or one in which the subjects do know which treatment was received, but the individuals measuring the response do not know which treatment was received.

▪ Experimental Units and Replication

An **experimental unit** is the smallest unit to which a treatment is applied. In the language of experimental design, treatments are assigned at random to experimental units, and replication means that each treatment is applied to more than one experimental unit.

Replication is necessary for randomization to be an effective way to create similar experimental groups and to get a sense of the variability in the values of the response for individuals that receive the same treatment. As we will see in later chapters (Chapters 9–15), this information is important because it enables us to use statistical methods to decide whether differences in the responses in different treatment groups can be attributed to the treatment received or if they can be explained by chance variation (the natural variability seen in the responses to a single treatment).

Be careful when designing an experiment to ensure that there is replication. For example, suppose that children in two third-grade classes are available to participate in an experiment to compare two different methods for teaching arithmetic. It might at first seem reasonable to select one class at random to use one method and then assign the other method to the remaining class. But what are the experimental units here? Treatments are randomly assigned to classes, and so classes are the experimental units. Because there are only two classes, with one assigned to each treatment, this is an experiment with no replication, even though there are many children in each class. We would *not* be able to determine whether there was a difference between the two methods based on data from this experiment, because we would have only one observation per treatment.

One last note on replication: Don't confuse replication in an experimental design with replicating an experiment. When investigators talk about replicating an experiment, they mean conducting a new experiment using the same experimental design as a previous experiment. Replicating an experiment is a way of confirming conclusions based on a previous experiment, but it does not eliminate the need for replication in each of the individual experiments themselves.

▪ Using Volunteers as Subjects in an Experiment

Although the use of volunteers in a study that involves collecting data through sampling is never a good idea, it is a common practice to use volunteers as subjects in an experiment. Even though the use of volunteers limits the researcher's ability to generalize to a larger population, random assignment of the volunteers to treatments should result in comparable groups, and so treatment effects can still be assessed.

▪ Exercises 2.49–2.58

2.49 Explain why some studies include both a control group and a placebo treatment. What additional comparisons are possible if both a control group and a placebo group are included?

2.50 Give an example of an experiment for each of the following:

a. Single-blind experiment with the subjects blinded

b. Single-blind experiment with the individuals measuring the response blinded

c. Double-blind experiment

d. An experiment that is not possible to blind

2.51 Explain why blinding is a reasonable strategy in many experiments.

2.52 A novel alternative medical treatment for heart attacks seeds the damaged heart muscle with cells

ı the patient's thigh muscle ("Doctors Mend ıaged Hearts with Cells from Muscles," *San Luis spo Tribune*, November 18, 2002). Doctor Dib ı the Arizona Heart Institute evaluated the ap- ıch on 16 patients with severe heart failure. The :le states that "ordinarily, the heart pushes out 'e than half its blood with each beat. Dib's pa- ts had such severe heart failure that their hearts ıped just 23 percent. After bypass surgery and injections, this improved to 36 percent, although 'as impossible to say how much, if any, of the new :ngth resulted from the extra cells."

Explain why it is not reasonable to generalize to population of all heart attack victims based on data from these 16 patients.

Explain why it is not possible to say whether any the observed improvement was due to the cell ections, based on the results of this study.

Describe a design for an experiment that would ow researchers to determine whether bypass :gery plus cell injections was more effective than pass surgery alone.

53 An article in the *San Luis Obispo Tribune* (Sep- ıber 7, 1999) described an experiment designed to vestigate the effect of creatine supplements on the velopment of muscle fibers. The article states that e researchers "looked at 19 men, all about 25 years age and similar in weight, lean body mass, and ıpacity to lift weights. Ten were given creatine —) grams a day for the first week, followed by 5 grams day for the rest of the study. The rest were given a ke preparation. No one was told what he was get- ng. All the men worked out under the guidance of ıe same trainer. The response variable measured 'as gain in fat-free mass (in percent)."

. What extraneous variables are identified in the iven statement, and what strategy did the research- rs use to deal with them?

. Do you think it was important that the men par- icipating in the experiment were not told whether hey were receiving creatine or the placebo? Explain.

. In a double-blind experiment, not only do the ubjects not know which treatment they are receiv- ng but also the person who measures the response 'ariable does not know which treatment the subjects ıre receiving. This particular experiment was not :onducted in a double-blind manner. Do you think it would have been a good idea to make this a double- ɔlind experiment? Explain.

2.54 An experiment to evaluate whether vitamins :an help prevent recurrence of blocked arteries in patients who have had surgery to clear blocked ar- teries was described in the article "Vitamins Found to Help Prevent Blocked Arteries" (Associated Press, September 1, 2002). The study involved 205 patients who were given either a treatment consisting of a combination of folic acid, vitamin B12, and vitamin B6 or a placebo for 6 months.

a. Explain why a placebo group was used in this experiment.

b. Explain why it would be important for the researchers to have assigned the 205 subjects to the two groups (vitamin and placebo) at random.

c. Do you think it is appropriate to generalize the results of this experiment to the population of all patients who have undergone surgery to clear blocked arteries? Explain.

2.55 Pismo Beach, California, has an annual clam festival that includes a clam chowder contest. Judges rate clam chowders from local restaurants, and the judging is done in such a way that the judges are not aware of which chowder is from which restaurant. One year, much to the dismay of the seafood restaurants on the waterfront, Denny's chowder was declared the winner! (When asked what the ingredients were, the cook at Denny's said he wasn't sure — he just had to add the right amount of nondairy creamer to the soup stock that he got from Denny's distribution center!)

a. Do you think that Denny's chowder would have won the contest if the judging had not been "blind"? Explain.

b. Although this was not an experiment, your answer to Part (a) helps to explain why those measuring the response in an experiment are often blinded. Using your answer in Part (a), explain why experiments are often blinded in this way.

2.56 The *San Luis Obispo Tribune* (May 7, 2002) reported that "a new analysis has found that in the majority of trials conducted by drug companies in recent decades, sugar pills have done as well as — or better than — antidepressants." What effect is being described here? What does this imply about the design of experiments with a goal of evaluating the effectiveness of a new medication?

2.57 Researchers at the University of Pennsylvania suggest that a nasal spray derived from pheromones (chemicals emitted by animals when they are trying to attract a mate) may be beneficial in relieving symptoms of premenstrual syndrome (PMS). Early tests indicated that the spray, called PH80, eases irritability and also reduces some physical symptoms (*Los Angeles Times*, January 17, 2003).

a. Describe how you might design an experiment using 100 female volunteers who suffer from PMS to determine whether the nasal spray reduces PMS symptoms.

▪ 2.7 Communicating and Interpreting the Results of Statistical Analyses

Statistical studies are conducted to allow investigators to answer questions about characteristics of some population of interest or about the effect of some treatment. Such questions are answered on the basis of data, and how the data are obtained determines the quality of information available and the type of conclusions that can be drawn. As a consequence, when describing a study you have conducted (or when evaluating a published study), you must consider how the data were collected.

The description of the data collection process should make it clear whether the study is an observational study or involves an experiment. For observational studies, some of the issues that should be addressed are:

1. What is the population of interest? What is the sampled population? Are these two populations the same? If the sampled population is only a subset of the population of interest, **undercoverage** limits our ability to generalize to the population of interest. For example, if the population of interest is all students at a particular university, but the sample is selected from only those students who choose to list their phone number in the campus directory, undercoverage may be a problem. We would need to think carefully about whether it is reasonable to consider the sample as representative of the population of all students at the university. For some purposes, this might be reasonable, but in other cases it may not be. **Overcoverage** results when the sampled population is actually larger than the population of interest. This would be the case if we were interested in the population of all high schools that offer Advanced Placement (AP) Statistics but sampled from a list of all schools that offered an AP class in any subject. Both undercoverage and overcoverage can be problematic.
2. How were the individuals or objects in the sample actually selected? A description of the sampling method facilitates making judgments about whether the sample can reasonably be viewed as representative of the population of interest.
3. What are potential sources of bias, and is it likely that any of these will have a substantial effect on the observed results? When describing an observational study, you should acknowledge that you are aware of potential sources of bias and explain any steps that were taken to minimize their effect. For example, in a mail survey, nonresponse can be a problem, but the sampling plan may seek to minimize its effect by offering incentives for participation and by following up one or more times with those who do not respond to the first request.

   A common misperception is that increasing the sample size is a way to reduce bias in observational studies, but this is not the case. For example, if measurement bias is present, as in the case of a scale that is not correctly calibrated and tends to weigh too high, taking 1000 measurements rather than 100 measurements cannot correct for the fact that the measured weights will be too large. Similarly, a larger sample size cannot compensate for response bias introduced by a poorly worded question.

For experiments, some of the issues that should be addressed are:

1. What is the role of randomization? All good experiments use random assignment as a means of coping with the effects of potentially confounding vari-

ables that cannot easily be directly controlled. When describing an experimental design, you should be clear about how random assignment (subjects to treatments, treatments to subjects, or treatments to trials) was incorporated into the design.

2. Were any extraneous variables directly controlled by holding them at fixed values throughout the experiment? If so, which ones and at which values?
3. Was blocking used? If so, how were the blocks created? If an experiment uses blocking to create groups of homogeneous experimental units, you should describe the criteria used to create the blocks and their rationale. For example, you might say something like "Subjects were divided into two blocks — those who exercise regularly and those who do not exercise regularly — because it was believed that exercise status might affect the responses to the diets."

   Because each treatment appears at least once in each block, the block size must be at least as large as the number of treatments. Ideally, the block sizes should be equal to the number of treatments, because this presumably would allow the experimenter to create small groups of extremely homogeneous experimental units. For example, in an experiment to compare two methods for teaching calculus to first-year college students, we may want to block on previous mathematics knowledge by using math SAT scores. If 100 students are available as subjects for this experiment, rather than creating two large groups (above-average math SAT score and below-average math SAT score), we might want to create 50 blocks of two students each, the first consisting of the two students with the highest math SAT scores, the second containing the two students with the next highest scores, and so on. We would then select one student in each block at random and assign that student to teaching method 1. The other student in the block would be assigned to teaching method 2.

▪ A Word to the Wise: Cautions and Limitations

It is a cardinal sin to begin collecting data before thinking carefully about research objectives and developing a research plan. Failure to do so may result in data that do not enable the researcher to answer key questions of interest or to generalize conclusions based on the data to the desired populations of interest.

Clearly defining the objectives at the outset enables the investigator to determine whether an experiment or an observational study is the best way to proceed. Watch out for the following:

1. Drawing a cause-and-effect conclusion from an observational study. Don't do this, and don't believe it when others do it!
2. Generalizing results of an experiment that uses volunteers as subjects to a larger population without a convincing argument that the group of volunteers can reasonably be considered a representative sample from the population.
3. Generalizing conclusions based on data from a sample to some population of interest. This is sometimes a sensible thing to do, but on other occasions it is not reasonable. Generalizing from a sample to a population is justified only when there is reason to believe that the sample is likely to be representative of the population. This would be the case if the sample was a random sample from the population and there were no major potential sources of bias. If the sample was not selected at random or if potential sources of bias were

present, these issues would have to be addressed before a judgment could be made regarding the appropriateness of generalizing the study results.

For example, the Associated Press (January 25, 2003) reported on the high cost of housing in California. The median home price was given for each of the 10 counties in California with the highest home prices. Although these 10 counties are a sample of the counties in California, they were not randomly selected and (because they are the 10 counties with the highest home prices) it would not be reasonable to generalize to all California counties based on data from this sample.

4. Generalizing conclusions based on an observational study that used voluntary response or convenience sampling to a larger population. This is almost never reasonable.

▪ Activity 2.1: Designing a Sampling Plan

Background: In this activity, you will work with a partner to develop a sampling plan.

Suppose that you would like to select a sample of 50 students at your school to learn something about how many hours per week, on average, students at your school spend engaged in a particular activity (such as studying, surfing the Internet, or watching TV).

1. Discuss with your partner whether you think it would be easy or difficult to obtain a simple random sample of students at your school and to obtain the desired information from all the students selected for the sample. Summarize your discussion by writing a few sentences explaining why you think it would be easy or difficult.

2. With your partner, decide how you might go about selecting a sample of 50 students from your school that, although it truly may not be a simple random sample, could be reasonably considered representative of the population of interest. Write a brief description of your sampling plan, and be sure to point out the aspects of your plan that you think make it reasonable to argue that it will be representative.

3. Explain your plan to another pair of students. Ask them to critique your plan, pointing out any potential flaws they see in it. Write a brief summary of the comments you received. Now reverse roles, and provide a critique of the plan devised by the other pair.

4. Based on the feedback you received in Step 3, would you modify your original sampling plan? If not, explain why this is not necessary. If so, describe how the plan would be modified.

▪ Activity 2.2: An Experiment to Test for the Stroop Effect

Background: In 1935, John Stroop published the results of his research into how people respond when presented with conflicting signals. Stroop noted that most people are able to read words quickly and that they cannot easily ignore them and focus on other attributes of a printed word, such as text color.

For example, consider the following list of words:

green blue red blue yellow red

It is easy to quickly read this list of words. It is also easy to read the words even if the words are printed in color, and even if the text color is different from the color of the word. For example, people can read the words in the list

green red blue yellow red

as quickly as they can read the list that isn't printed in color.

However, Stroop found that if people are asked to name the text colors of the words in the list (red, yellow, blue, green, red, green), it takes them longer. Psychologists believe that this is because the reader has to inhibit a natural response (reading the word) and produce a different response (naming the color of the text).

If Stroop is correct, people should be able to name colors more quickly if they do not have to inhibit the word response, as would be the case if they were shown the following:

esign an experiment to compare times to iden-olors when they appear as text to times to iden-olors when there is no need to inhibit a word nse. Be sure to indicate how randomization is porated into your design. What is your response ble? How will you measure it? How many sub- will you use in your experiment, and how will be chosen?

2. When you are satisfied with your experimental design, carry out the experiment. You will need to construct your list of colored words and a corresponding list of colored bars to use in the experiment. You will also need to think about how you will implement your randomization scheme.

3. Summarize the resulting data in a brief report that explains whether your findings are consistent with the Stroop effect.

Summary of Key Concepts and Formulas

n or Formula	Comment
ervational study	A study that observes characteristics of an existing population.
ple random sample ze n	A sample selected in a way that gives every different sample of size n an equal chance of being selected.
tified sampling	Dividing a population into subgroups (strata) and then taking a separate random sample from each stratum.
ster sampling	Dividing a population into subgroups (clusters) and forming a sample by randomly selecting clusters and including all individuals or objects in the selected clusters in the sample.
k systematic sampling	A sample selected from an ordered arrangement of a population by choosing a starting point at random from the first k individuals on the list and then selecting every kth individual thereafter.
nfounding variable	A variable that is related both to group membership and to the response variable.
asurement or onse bias	The tendency for a sample to differ from the population because the method of observation tends to produce values that differ from the true value.
ection bias	The tendency for a sample to differ from the population because of systematic exclusion of some part of the population.
nresponse bias	The tendency for a sample to differ from the population because measurements are not obtained from all individuals selected for inclusion in the sample.
periment	A procedure for investigating the effect of *experimental conditions* (which are manipulated by the experimenter) on a *response variable*.
eatments	The experimental conditions imposed by the experimenter.
traneous factor	A variable that is not of interest in the current study but is thought to affect the response variable.
ect control	Holding extraneous factors constant so that their effects are not confounded with those of the experimental conditions.
ocking	Using extraneous factors to create experimental groups that are similar with respect to those factors, thereby filtering out their effect.
ndomization	Random assignment of experimental units to treatments or of treatments to trials.