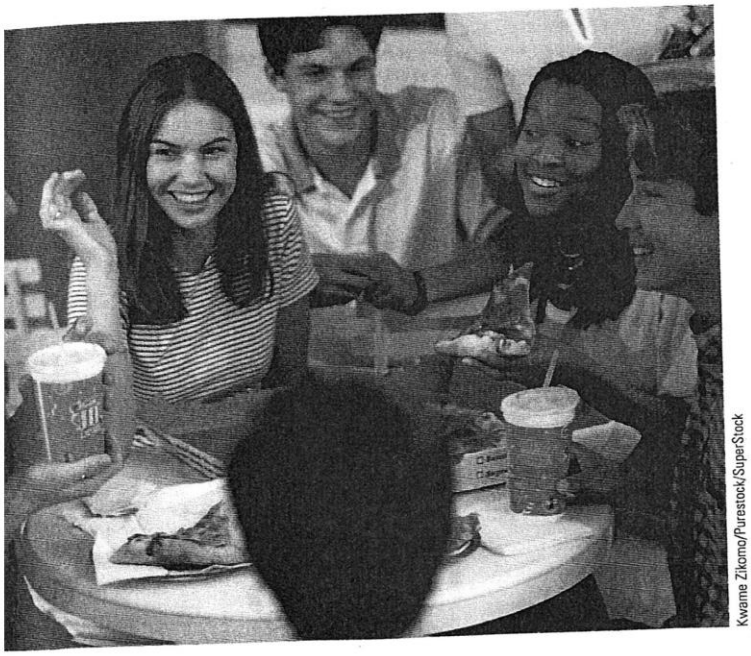
CHAPTER 2

Collecting Data Sensibly



Kwame Zikomo/Purestock/SuperStock

Data and conclusions from data are everywhere—in newspapers, magazines, online resources, and professional publications. But should you believe what you read? For example, should you drink hot chocolate to improve your memory? Will eating beef make you happier? Will thinking positive thoughts add 7 years to your life? These are just three of many claims made in one issue of *Woman's World* (December 23, 2013), a magazine with over 1.5 million readers. The magazine suggests that these claims are based on research studies, but how reliable are these studies? Are the conclusions drawn reasonable, and do they apply to you? These are important questions.

A primary goal of statistical studies is to collect data that can then be used to make informed decisions. It should come as no surprise that the ability to make good decisions depends on the quality of the information available.

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 consider two types of statistical studies and then focus on two widely used methods of data collection: sampling and experimentation.

Chapter 2: Learning Objectives

STUDENTS WILL UNDERSTAND:

- that the types of conclusions that can be drawn from data depend on the way data were collected.
- that bias may be present when data are collected from a sample.
- why random selection is an important component of a sampling plan.
- why random assignment is important when collecting data in an experiment.
- the purposes of a control group and blinding in an experiment.

STUDENTS WILL BE ABLE TO:

- distinguish between an observational study and an experiment.
- distinguish between selection bias, measurement or response bias, and nonresponse bias.
- select a simple random sample from a given population.

- distinguish between simple random sampling, stratified random sampling, cluster sampling, systematic sampling, and convenience sampling.
- describe a procedure for randomly assigning subjects to treatments in an experiment.
- design a completely randomized experiment.
- design a randomized block experiment.

2.1 Statistical Studies: Observation and Experimentation

On September 25, 2009, results from a study of the relationship between spanking and IQ were reported by a number of different news media. Some of the headlines that appeared that day were:

"Spanking lowers a child's IQ" (*Los Angeles Times*)

"Do you spank? Studies indicate it could lower your kid's IQ" (*SciGuy, Houston Chronicle*)

"Spanking can lower IQ" (NBC4i, Columbus, Ohio)

"Smacking hits kids' IQ" (*newscientist.com*)

In the study that these headlines refer to, the investigators followed 806 kids age 2 to 4 and 704 kids age 5 to 9 for 4 years. IQ was measured at the beginning of the study and again 4 years later. The researchers found that at the end of the study, the average IQ of kids in the younger group who were not spanked was 5 points higher than that of kids who were spanked. For the older group, the average IQ of kids who were not spanked was 2.8 points higher.

These headlines all imply that spanking was the cause of the observed difference in IQ. Is this conclusion reasonable? The answer depends in a critical way on the study design. After considering some important aspects of study design, we'll return to these headlines and decide if they are appropriate.

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 a single existing population or in comparing two or more well-defined populations. To accomplish this, we select a sample from each population under consideration and use the sample information to gain insight into characteristics of those populations.

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. These are examples of studies that are *observational* in nature. In these studies, 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 **observational studies**, it is important to obtain samples that are representative of the corresponding populations.

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 hours to 6 hours per

week. To answer such questions, 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-hour lab and 6-hour lab). In an **experiment**, the researcher manipulates one or more explanatory variables, also sometimes called factors, to create the experimental conditions.

DEFINITION

Observational study: A study in which the investigator observes characteristics of a sample selected from 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. In a well-designed observational study, the sample is selected in a way that is designed to produce a sample that is respresentative of the population.

Experiment: A study in which the investigator observes how a response variable behaves when one or more explanatory variables, also called factors, are manipulated. The usual goal of an experiment is to determine the effect of the manipulated explanatory variables (factors) on the response variable. In a well-designed experiment, the composition of the groups that will be exposed to different experimental conditions is determined by random assignment.

The type of conclusion that can be drawn from a statistical study depends on how the study was conducted. 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 drawing conclusions based on data 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 clear cause-and-effect conclusions because we cannot rule out the possibility that the observed effect is due to some variable other than the explanatory variable being studied. Such variables are called confounding variables.

DEFINITION

Confounding variable: A variable that is related to both how the experimental groups were formed and the response variable of interest.

Consider the role of confounding variables in the following three studies:

- The article "Panel Can't Determine the Value of Daily Vitamins" (*San Luis Obispo Tribune*, July 1, 2003) summarized conclusions from a government advisory panel that investigated the benefits of vitamin use. The panel looked at a large number of studies on vitamin use and concluded that the results were "inadequate or conflicting." A major concern was that many of the studies were observational in nature and the panel worried that people who take vitamins might be healthier just because they tend to take better care of themselves in general. This potential confounding variable prevented the panel from concluding that taking vitamins is the *cause* of observed better health among those who take vitamins.
- Studies have shown that people over age 65 who get a flu shot are less likely to die from a flu-related illness during the following year than those who do not get a flu shot. However, recent research has shown that people over age 65 who get

a flu shot are also less likely to die from *any* cause during the following year than those who don't get a flu shot (*International Journal of Epidemiology*, December 21, 2005). This has led to the speculation that those over age 65 who get flu shots are healthier as a group than those who do not get flu shots. If this is the case, observational studies that compare two groups—those who get flu shots and those who do not—may overestimate the effectiveness of the flu vaccine because general health differs in the two groups. General health is a possible confounding variable in such studies.

- The article "Heartfelt Thanks to Fido" (*San Luis Obispo Tribune*, July 5, 2003) summarized a study that appeared in the *American Journal of Cardiology* (March 15, 2003). In this study researchers measured heart rate variability (a measure of the heart's ability to handle stress) in patients who had recovered from a heart attack. They found that heart rate variability was higher (which is good and means the heart can handle stress better) for those who owned a dog than for those who did not. Should someone who suffers a heart attack immediately go out and get a dog? Well, maybe not yet. The American Heart Association recommends additional studies to determine if the improved heart rate variability is attributable to dog ownership or due to the fact that dog owners get more exercise. If in fact dog owners do tend to get more exercise than nonowners, level of exercise is a confounding variable that would prevent us from concluding that owning a dog is the *cause* of improved heart rate variability.

Each of the three studies described above illustrates why potential confounding variables make it unreasonable to draw a cause-and-effect conclusion from an observational study.

Let's return to the study on spanking and IQ described at the beginning of this section. Is this study an observational study or an experiment? Two groups were compared (children who were spanked and children who were not spanked), but the researchers did not randomly assign children to the spanking or no-spanking groups. The study is observational, and so cause-and-effect conclusions such as "spanking lowers IQ" are not justified based on the observed data. What we can say is that there is evidence that, as a group, children who are spanked tend to have a lower IQ than children who are not spanked. What we cannot say is that spanking is the *cause* of the lower IQ. It is possible that other variables—such as home or school environment, socio-economic status, or parents' education—are related to both IQ and whether or not a child was spanked. These are examples of possible confounding variables.

Fortunately, not everyone made the same mistake as the writers of the headlines given earlier in this section. Some examples of headlines that got it right are:

"Lower IQ's measured in spanked children" (*world-science.net*)

"Children who get spanked have lower IQs" (*livescience.com*)

"Research suggests an association between spanking and lower IQ in children" (*CBSnews.com*)

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.

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. Both observational studies and experiments must be carefully designed if the resulting data are to be useful. The common sampling procedures used in observational studies are considered in Section 2.2. In Sections 2.3 and 2.4, we consider experimentation and explore what constitutes good practice in the design of simple experiments.

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
• 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

EXERCISES 2.1 - 2.12

2.1 The article **"How Dangerous Is a Day in the Hospital?"** (*Medical Care* [2011]: 1068–1075) describes a study to determine if the risk of an infection is related to the length of a hospital stay. The researchers looked at a large number of hospitalized patients and compared the proportion who got an infection for two groups of patients—those who were hospitalized overnight and those who were hospitalized for more than one night. Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.2 The authors of the paper **"Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course"** (*Teaching in Psychology* [2007]: 245–247) carried out a study to see if events unrelated to an undergraduate course could affect student evaluations. Students enrolled in statistics courses taught by the same instructor participated in the study. All students attended the same lectures and one of six discussion sections that met once a week. At the end of the course, the researchers chose three of the discussion sections to be the "chocolate group." Students in these three sections were offered chocolate prior to having them fill out course evaluations. Students in the other three sections were not offered chocolate.

The researchers concluded that "Overall, students offered chocolate gave more positive evaluations than students not offered chocolate." Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.3 The article **"Why We Fall for This"** (*AARP Magazine*, May/June 2011) described a study in which a business professor divided his class into two groups. He showed students a mug and then asked students in one of the groups how much they would pay for the mug. Students in the other group were asked how much they would sell the mug for if it belonged to them. Surprisingly, the average value assigned to the mug was quite different for the two groups! Indicate whether the study is an observational study or an experiment. Give a brief explanation for your choice.

2.4 ▼ The article **"Television's Value to Kids: It's All in How They Use It"** (*Seattle Times*, July 6, 2005) described a study in which researchers analyzed standardized test results and television viewing

habits of 1700 children. They found that children who averaged more than 2 hours of television viewing per day when they were younger than 3 tended to score lower on measures of reading ability and short-term memory.

a. Is the study described an observational study or an experiment?

b. Is it reasonable to conclude that watching 2 or more hours of television is the cause of lower reading scores? Explain. (Hint: Look at Table 2.1.)

2.5 The article **"Acupuncture for Bad Backs: Even Sham Therapy Works"** (*Time*, **May 12, 2009**) summarized a study conducted by researchers at the Group Health Center for Health Studies in Seattle. In this study, 638 adults with back pain were randomly assigned to one of four groups. People in group 1 received the usual care for back pain. People in group 2 received acupuncture at a set of points tailored specifically for each individual. People in group 3 received acupuncture at a standard set of points typically used in the treatment of back pain. Those in group 4 received fake acupuncture—they were poked with a toothpick at the same set of points chosen for the people in group 3!

Two notable conclusions from the study were: (1) patients receiving real or fake acupuncture experienced a greater reduction in pain than those receiving usual care; and (2) there was no significant difference in pain reduction for those who received acupuncture (at individualized or the standard set of points) and those who received fake acupuncture toothpick pokes.

a. Is this study an observational study or an experiment? Explain.

b. Is it reasonable to conclude that receiving either real or fake acupuncture was the cause of the observed reduction in pain in those groups compared to the usual care group? What aspect of this study supports your answer? (Hint: Look at Table 2.1.)

2.6 The article **"Display of Health Risk Behaviors on MySpace by Adolescents"** (*Archives of Pediatrics and Adolescent Medicine* **[2009]: 27–34**) described a study in which researchers looked at a random sample of 500 publicly accessible MySpace web profiles posted by 18-year-olds. The content of each profile was analyzed. One of the conclusions reported was that displaying sport or hobby involvement was associated with decreased references to risky behavior (sexual references or references to substance abuse or violence).

a. Is the study described an observational study or an experiment?

b. Is it reasonable to generalize the stated conclusion to all 18-year-olds with a publicly accessible MySpace web profile? What aspect of the study supports your answer?

c. Not all MySpace users have a publicly accessible profile. Is it reasonable to generalize the stated conclusion to all 18-year-old MySpace users? Explain.

d. Is it reasonable to generalize the stated conclusion to all MySpace users with a publicly accessible profile? Explain.

2.7 Can choosing the right music make wine taste better? This question was investigated by a researcher at a university in Edinburgh (www.decanter.com/news). Each of 250 volunteers was assigned at random to one of five rooms where they were asked to taste and rate a glass of wine. In one of the rooms, no music was playing and a different style of music was playing in each of the other four rooms. The researchers concluded that cabernet sauvignon is perceived as being richer and more robust when bold music is played than when no music is heard.

a. Is the study described an observational study or an experiment?

b. Can a case be made for the researcher's conclusion that the music played was the cause for the higher rating? Explain.

2.8 **"Fruit Juice May Be Fueling Pudgy Preschoolers, Study Says"** is the title of an article that appeared in the *San Luis Obispo Tribune* **(February 27, 2005)**. This article describes a study that found that for 3- and 4-year-olds, drinking something sweet once or twice a day doubled the risk of being seriously overweight one year later. The authors of the study state

> Total energy may be a confounder if consumption of sweet drinks is a marker for other dietary factors associated with overweight (***Pediatrics*, November 2005**).

Give an example of a dietary factor that might be one of the potentially confounding variables the study authors are worried about.

2.9 The article **"Americans are 'Getting the Wrong Idea' on Alcohol and Health"** (**Associated Press, April 19, 2005**) reported that observational studies in recent years that have concluded that moderate drinking is associated with a reduction in the risk of heart disease may be misleading. The article refers to a study conducted by the Centers for Disease Control and Prevention that showed that moderate drinkers, as a group, tended to

be better educated, wealthier, and more active than nondrinkers.

Explain why the existence of these potentially confounding variables prevents drawing the conclusion that moderate drinking is the cause of reduced risk of heart disease.

2.10 Based on a survey conducted on the eDiets.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.

2.11 ▼ A survey of affluent Americans (those with incomes of $75,000 or more) indicated that 57% would rather have more time than more money (*USA Today*, January 29, 2003).

a. What condition on how the data were collected would make the generalization from the sample to the population of affluent Americans reasonable?

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

2.12 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 indicates that a greater percentage of Southerners have high blood pressure than do people in any other region of the United States (see CNN.com web site article of January 6, 2000, titled "High Blood Pressure Greater Risk in U.S. South, Study Says"). 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 why we cannot conclude that living in the South causes high blood pressure.

Bold exercises answered in back ● Data set available online ▼ Video Solution available

2.2 Sampling

Many studies are conducted in order to generalize the results of the study to the corresponding population. In this case, it is important that the sample be representative of the population. To be reasonably sure of this, we 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 is not possible 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. When we attempt to generalize from this convenience sample, the difficulty is that it is not clear how these factors (and others that we might not be aware of) affect any conclusions 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.

There are many reasons for selecting a sample rather than obtaining information from an entire population (a **census**). Sometimes the process of measuring the characteristics of interest is destructive, as with measuring the lifetime of flashlight batteries or the sugar content of oranges. It would be foolish to study the entire population in situations like these. But the most common reason for selecting a sample is limited resources.

Restrictions on available time or money usually make it impossible to collect data from an entire population.

Bias in Sampling

Bias in sampling is the tendency for samples 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 (sometimes also called undercoverage) is introduced when the way the sample is selected systematically excludes some part of the population of interest. 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 tend to 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 virtually 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 for the population of interest.

Selection bias also occurs if only volunteers or self-selected individuals are used in a study, because those who choose to participate (for example, in a call-in telephone poll) may 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 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 not to 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). Response bias is typically used to describe systematic deviations from the true value when people provide answers to survey questions.

Nonresponse bias occurs when responses are not 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, and 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 (available online).

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 if the method of selecting the sample is flawed or if the nonresponse rate remains high.

Potential sources of bias are illustrated in the following examples.

EXAMPLE 2.1 Are Cell Phone Users Different?

Understand the context)

Many surveys are conducted by telephone and participants are often selected from phone books that include only landline telephones. For many years, it was thought that this was not a serious problem because most cell phone users also had a landline phone and so they still had a chance of being included in the survey. But the number of people with cell phones only is growing, and this trend is a concern for survey organizations.

The article **"Omitting Cell Phone Users May Affect Polls" (*Associated Press*, September 25, 2008)** described a study that examined whether people who only have a cell phone are different from those who have landline phones. One finding from the study was that for people under the age of 30 with only a cell phone, 28% were Republicans compared to 36% of landline users. This suggests that researchers who use telephone surveys need to worry about how selection bias might influence the ability to generalize the results of a survey if only landlines are used. ■

EXAMPLE 2.2 Think Before You Order That Burger!

Understand the context)

The article **"What People Buy from Fast-Food Restaurants: Caloric Content and Menu Item Selection" (*Obesity* [2009]: 1369–1374)** reported that the average number of calories consumed at lunch in New York City fast-food restaurants was 827. The researchers selected 267 fast-food locations at random. The paper states that at each of these locations "adult customers were approached as they entered the restaurant and asked to provide their food receipt when exiting and to complete a brief survey."

Approaching customers as they entered the restaurant and before they ordered may have influenced what they purchased. This introduces the potential for response bias. In addition, some people chose not to participate when approached. If those who chose not to participate differed from those who did participate, the researchers also need to be concerned about nonresponse bias. Both of these potential sources of bias limit the researchers' ability to generalize conclusions based on data from this study. ■

Random Sampling

Most of the methods introduced in this text are based on the idea of random selection. The most straightforward sampling method 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, the method used to select the sample must ensure that each of the many different subsets of 10 employees must be equally likely to be 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 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, which 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 just described, $n = 10$ becasue 10 employees were to be selected.

DEFINITION

Simple random sample of size *n*: A sample that is selected from a population in a way that ensures that every different possible sample of size n 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 4 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 10 students—for example, a sample of 7 females and 3 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 stratified sampling, to be discussed in more detail shortly). But it does not produce a simple random sample. When this is the case, it is sometimes necessary to use different methods when generalizing results from the sample to the population. For this reason, the choice of sampling method is an important consideration that must be considered when a method is chosen for analyzing data resulting from such a sampling method.

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 different but identical slips of paper. 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 relatively small populations.

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. A table of random digits or a random number generator can then be used to select the sample. A random number generator is a procedure that produces a sequence of numbers that satisfies 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 A, Table 1.

For example, suppose a list containing the names of the 427 customers who purchased a new car during 2014 at a large dealership is available. The owner of the dealership wants to interview a sample of these customers to learn about customer satisfaction. She plans to select a simple random sample of 20 customers. Because it would be tedious to write all 427 names on slips of paper, random numbers can be used to select the sample. To do this, we can use three-digit numbers, starting with 001 and ending with 427, to represent the individuals on the list.

The random digits from rows 6 and 7 of Appendix A, Table 1 are shown here:

09387679956256584264
41010220475119479751

We can use blocks of three digits from this list (underlined in the lists above) to identify the individuals who should be included in the sample. The first block of three digits is 093, so the 93rd person on the list will be included in the sample. The next five blocks of three digits (876, 799, 562, 565, and 842) do not correspond to anyone on the list, so we ignore them. The next block that corresponds to a person on the list is 410, so that person is included in the sample. This process would continue until 20 people have been selected for the sample. We would ignore any three-digit repeats since any particular person should only be selected once for the sample.

Another way to select the sample would be to use computer software or a graphing calculator to generate 20 random numbers. For example, Minitab produced the following numbers when 20 random numbers between 1 and 427 were requested.

289	67	29	26	205	214	422	31	233	98
10	203	346	186	232	410	43	293	25	371

These numbers could be used to determine which 20 customers to include in the sample.

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. 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**.

DEFINITION

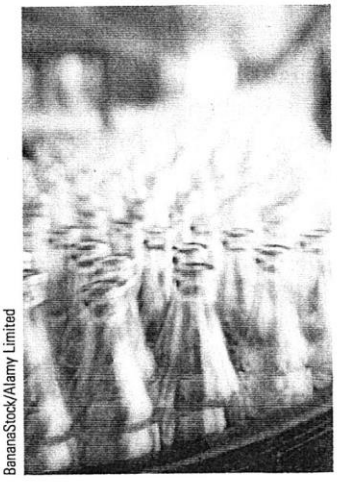
Sampling without replacement: Once an individual from the population is selected for inclusion in the sample, it may not be selected again in the sampling process. A sample selected without replacement includes n distinct individuals from the population.

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 methods can be viewed as equivalent if the sample size is less than 10% of the population size.

Step-by-step technology instructions available online

Understand the context)

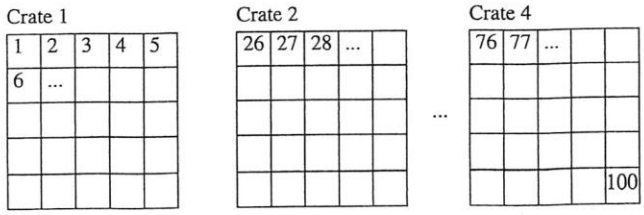


BananaStock/Alamy Limited

Formulate a plan)

EXAMPLE 2.3 Selecting a Random Sample of Glass Soda Bottles

Breaking strength is an important characteristic of glass soda bottles. 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 in each crate, starting with the top row of crate 1, as pictured:



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 the 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. ■

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, it does allow us to assess the risk of an unrepresentative sample. It is the ability to quantify this risk that will enable us to generalize with confidence from a random sample to the corresponding population.

An Important 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 cannot 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 is 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 five different random samples from the population, ranging in sample size from $n = 50$ to $n = 1000$.

Notice that each of 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 math 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 allows us to be confident that the resulting sample adequately reflects the population, even when the sample consists of only a small fraction of the population.

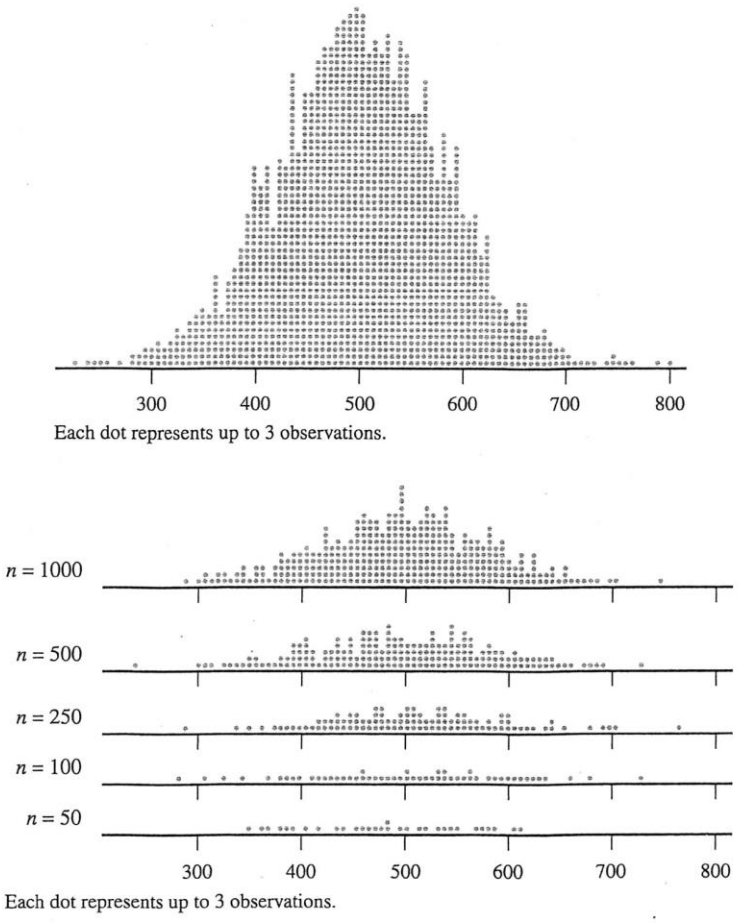


FIGURE 2.1
(a) Dotplot of math SAT scores for the entire population.
(b) Dotplots of math SAT scores for random samples of sizes 50, 100, 250, 500, and 1000.

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, and sometimes even 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 simple 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 city 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 random simple sample from the population of all doctors, the researcher could take four separate simple 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** and each individual subgroup is called a stratum (the singular of strata). Stratified sampling entails selecting a separate simple random sample from each stratum. Stratified sampling can be used instead of simple 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 produce relatively accurate estimates of 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. The individual GPAs of these students are all quite similar (a homogeneous group), and even a sample of three or four individuals from this subpopulation should be representative. On the other hand, producing a reasonably accurate estimate of the average GPA of *all* seniors at the university, a much more diverse group of GPAs, is a more difficult task. This means that 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 a stratified random sample can produce more accurate estimates of population characteristics than a simple random sample of the same size.

Cluster Sampling

Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves. **Cluster sampling** involves dividing the population of interest into nonoverlapping subgroups, called **clusters**. Clusters are then selected at random, and then *all* individuals in the selected clusters are included in the sample.

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 about 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. Then a survey could be administered to all students in the selected homerooms at the same time—certainly easier logistically than randomly selecting 75 individual seniors and then administering the survey to these students.

Because whole clusters are selected, the ideal situation for cluster sampling is 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 are 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 subgroup, whereas in cluster sampling, we include only selected whole clusters 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 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 (for example, $k = 50$ 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 $k = 20$ 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 (because $900/100 = 9$).

As long as there are no repeating patterns in the population sequence, systematic sampling works reasonably well. However, if there are such patterns, systematic sampling can result in an unrepresentative sample. For example, suppose that workers at the entry station of a state park have recorded the number of visitors to the park each day for the past 10 years. In a 1 in 70 systematic sample of days from this list, we would pick one of the first 70 days at random and then every 70th day after that. But if the first day selected happened to be a Wednesday, *every* day selected in the entire sample would also be a Wednesday (because there are 7 days a week and 70 is a multiple of 7). It is unlikely that such a sample would be representative of the entire collection of days. The number of visitors is likely to be higher on weekend days, and no Saturdays or Sundays would be included in the 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 2.13 - 2.32

2.13 A New York psychologist recommends that if you feel the need to check your e-mail in the middle of a movie or if you sleep with your cell phone next to your bed, it might be time to "power off" (***AARP Bulletin*, September 2010**). Suppose that you want to learn about the proportion of students at your college who would feel the need to check e-mail during the middle of a movie and that you have access to a list of all students enrolled at your college. Describe how you would use this list to select a simple random sample of 100 students.

2.14 As part of a curriculum review, the psychology department would like to select a simple random sample of 20 of last year's 140 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.

2.15 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.

2.16 The article **"Bicyclists and Other Cyclists"** (***Annals of Emergency Medicine* [2010]: 426**) reported that in 2008, there were 716 bicyclists killed on public roadways in the United States, and that the average age of the cyclists killed was 41 years. These figures were based on an analysis of the records of all traffic-related deaths of bicyclists on U.S. public roadways (this information is kept by the National Highway Traffic Safety Administration).

a. Does the group of 716 bicycle fatalities represent a census or a sample of the 2008 bicycle fatalities?

b. If the population of interest is 2008 bicycle traffic fatalities, is the given average age of 41 years a number that describes a sample or a number that describes the population?

2.17 The article **"Teenage Physical Activity Reduces Risk of Cognitive Impairment in Later Life"** (***Journal of the American Geriatrics Society*** **[2010]**) describes a study of more than 9000 women from Maryland, Minnesota, Oregon, and Pennsylvania. The women were asked about their physical activity as teenagers and at ages 30 and 50. A press release about this study (www.wiley.com) generalized the results of this study to all American women. In the press release, the researcher who conducted the study is quoted as saying

> Our study shows that women who are regularly physically active at any age have lower risk of cognitive impairment than those who are inactive but that being physically active at teenage is most important in preventing cognitive impairment.

Answer the following four questions for this observational study. (Hint: Reviewing Examples 2.1 and 2.2 might be helpful.)

a. What is the population of interest?

b. Was the sample selected in a reasonable way?

c. Is the sample likely to be representative of the population of interest?

d. Are there any obvious sources of bias?

2.18 ▼ For each of the situations described, state whether the sampling procedure is simple random sampling, stratified random sampling, cluster sampling, systematic sampling, or convenience sampling.

a. All first-year students at a university are enrolled in one of 30 sections of a seminar course. To select a sample of freshmen at this university, a researcher selects four sections of the seminar course at random from the 30 sections and all students in the four 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.

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

a. Describe a procedure for selecting a stratified random sample that uses full-time and part-time students as the two strata and that includes 10 students from each stratum.

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

2.20 Briefly explain why it is advisable to avoid the use of convenience samples.

2.21 A sample of pages from this 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.

2.22 In 2000, the chairman of a California ballot initiative campaign to add "none of the above" to the list

students? Address at least two possible sources of bias in your answer.

2.29 The financial aid advisor of a university plans to use a stratified random sample to estimate the average amount of money that students spend on textbooks each term. For each of the following proposed stratification schemes, discuss whether it would be worthwhile to stratify the university students in this manner. (Hint: Remember that it is desirable to create strata that are homogeneous.)

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.

2.30 Suppose that you were asked to help design a survey of adult city residents in order to estimate the proportion who would support a sales tax increase. The plan is to use a stratified random sample, and three stratification schemes have been proposed.

Scheme 1: Stratify adult residents into four strata based on the first letter of their last name (A–G, H–N, O–T, U–Z).

Scheme 2: Stratify adult residents into three strata: college students, nonstudents who work full time, nonstudents who do not work full time.

Scheme 3: Stratify adult residents into five strata by randomly assigning residents into one of the five strata.

Which of the three stratification schemes would be best in this situation? Explain.

2.31 The article "High Levels of Mercury Are Found in Californians" (*Los Angeles Times*, February 9, 2006) describes a study in which hair samples were tested for mercury. The hair samples were obtained from more than 6000 people who voluntarily sent hair samples to researchers at Greenpeace and The Sierra Club. The researchers found that nearly one-third of those tested had mercury levels that exceeded the concentration thought to be safe. Is it reasonable to generalize this result to the larger population of U.S. adults? Explain why or why not.

2.32 ▼ Whether or not to continue a Mardi Gras Parade through downtown San Luis Obispo, CA, is a hotly debated topic. The parade is popular with students and many residents, but some celebrations have led to complaints and a call to eliminate the parade. The local newspaper conducted online and telephone surveys of its readers and was surprised by the results. The survey web site received more than 400 responses, with more than 60% favoring continuing the parade, while the telephone response line received more than 120 calls, with more than 90% favoring banning the parade (*San Luis Obispo Tribune*, March 3, 2004). What factors may have contributed to these very different results?

Bold exercises answered in back ● Data set available online ▼ Video Solution available

2.3 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 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. Experiments provide a way to collect data to answer these types of questions.

DEFINITION

Experiment: A study in which one or more explanatory variables are manipulated in order to observe the effect on a response variable.

Explanatory variables: Those variables that have values that are controlled by the experimenter. Explanatory variables are also called **factors**.

Response variable: A variable that is thought to be related to the explanatory variable in an experiment. It is measured as part of the experiment, but it is not controlled by the experimenter.

Experimental condition: Any particular combination of values for the explanatory variables. Experimental conditions are also called **treatments**.

Suppose we are interested in determining the effect of room temperature on performance on a first-year calculus exam. In this case, the explanatory variable is room temperature (it can be manipulated by the experimenter). The response variable is exam performance (the variable that is not controlled by the experimenter and that will be measured).

In general, we can identify the explanatory variables and the response variable easily if we can describe the purpose of the experiment in the following terms:

The purpose is to assess the effect of _____ on _____.
explanatory variable — response variable

Let's return to the example of an experiment to assess the effect of room temperature on exam performance. We might decide to use two room temperature settings, 65° and 75°. This would result in an experiment with two experimental conditions (or equivalently, two treatments) corresponding to the two temperature settings.

Suppose that there are 10 sections of first-semester calculus that have agreed to participate in our study. We might design an experiment in this way: Set the room temperature (in degrees Fahrenheit) to 65° in five of the rooms and to 75° in the other five 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?

Based on the information given, 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? 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 other possible explanations for any observed differences in the response variable.

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

DEFINITION

Extraneous variable: A variable that is not one of the explanatory variables in the study but is thought to affect the response variable.

A well-designed experiment copes with the potential effects of extraneous variables by using **random assignment** to experimental conditions and sometimes also by incorporating direct control and/or blocking into the design of the experiment. Each of these strategies—random assignment, direct control, and blocking—is described in the paragraphs that follow.

A researcher can **directly control** some extraneous variables. In the calculus test example, the textbook used is an extraneous variable because part of the differences in test results might be attributed to this variable. We could control this variable directly, by requiring that all sections use the same textbook. Then any observed differences in test scores between temperature groups could not be explained by the use of different textbooks. The

extraneous variable *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 variables can be filtered out by a process known as **blocking**. Extraneous variables that are addressed through blocking are called *blocking variables*. An investigator using blocking creates groups (called blocks) that are similar with respect to blocking variables. Then all treatments are tried in each block. In our example, we might use *instructor* as a blocking variable. If five 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 extraneous variable *instructor* can be ruled out as a possible explanation, because all five instructors' students were present in each temperature group. (Had we controlled the instructor variable by choosing to have only one instructor, that would be an example of direct control. Of course we can't directly control both time of day and instructor.)

If one instructor taught all the 65° sections and another taught all the 75° sections, we would be unable to distinguish the effect of temperature from the effect of the instructor. In this situation, the two variables (temperature and instructor) are said to be **confounded**.

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

If an extraneous variable is confounded with the explanatory variables (which define the treatments), it is not possible to draw an unambiguous conclusion about the effect of the treatment on the response. Both direct control and blocking are effective in ensuring that the controlled variables and blocking variables are not confounded with the variables that define the treatments.

We can directly control some extraneous variables by holding them constant, and we can use blocking to create groups that are similar to essentially filter out the effect of other extraneous variables. But what about variables, such as student ability in our calculus test example, which cannot be controlled by the experimenter and which would be difficult to use as blocking variables? These extraneous variables are handled by the use of **random assignment** to experimental groups.

Random assignment ensures that our experiment does not systematically favor one experimental condition over any other and attempts to create experimental groups that are as much alike as possible. For example, if the students requesting calculus could be assigned to one of the ten available sections using a random mechanism, we would expect the resulting groups to be similar with respect to student ability as well as with respect to other extraneous variables that are not directly controlled or used as a basis for blocking.

Notice that random assignment in an experiment is different from random selection of subjects. The ideal situation would be to have both random selection of subjects and random assignment of subjects to experimental conditions, as this would allow conclusions from the experiment to be generalized to a larger population.

For many experiments the random selection of subjects is not possible. As long as subjects are assigned at random to experimental conditions, it is still possible to assess treatment effects.

To get a sense of how random assignment tends to create similar groups, suppose that 50 college freshmen 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.

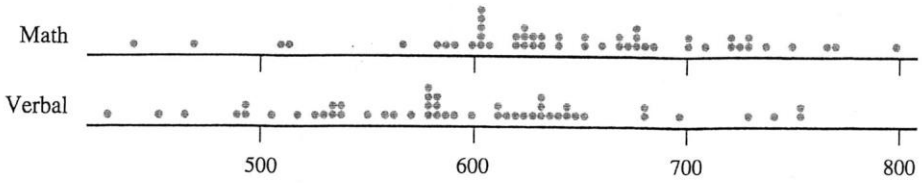


FIGURE 2.2
Dotplots of math and verbal SAT scores for 50 freshmen.

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 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 math 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 verbal 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.

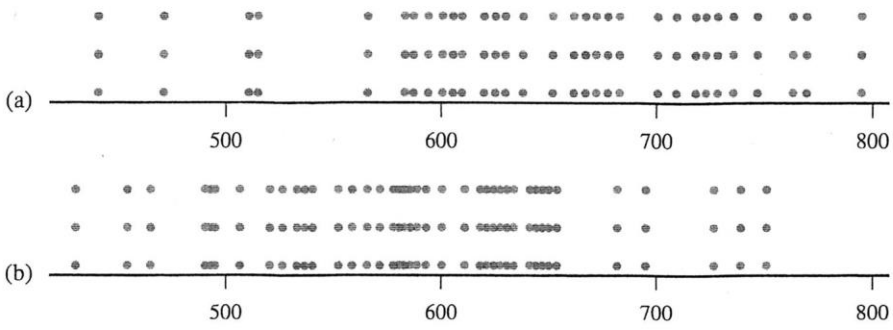


FIGURE 2.3
Dotplots for three different random assignments to two groups, one shown in orange and one shown in blue:
(a) math SAT score;
(b) verbal SAT score.

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 random assignment is a part of all well-designed experiments.

Not all experiments require the use of human subjects. For example, a researcher interested in comparing the effect of three different gasoline additives on gas mileage might conduct an experiment 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—10, for example—with each additive.

The experiment just described can be viewed as consisting of a sequence of trials. Because a number of extraneous variables (such as variations in environmental conditions like wind speed or humidity and small variations in the condition of the car) might have an effect on gas mileage, 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. A better approach would be to randomly assign additive 1 to 10 of the 30 planned trials, and then randomly assign additive 2 to 10