

# Designing Studies

## *Can Magnets Help Reduce Pain?*

Early research showed that magnetic fields affected living tissue in humans. Some doctors have begun to use magnets to treat patients with chronic pain. Scientists wondered whether this type of therapy really worked. They designed a study to find out.

Fifty patients with chronic pain were recruited for the study. A doctor identified a painful site on each patient and asked him or her to rate the pain on a scale from 0 (mild pain) to 10 (severe pain). Then, the doctor selected a sealed envelope containing a magnet from a box that contained both active and inactive magnets. That way, neither the doctor nor the patient knew which type of magnet was being used. The chosen magnet was applied to the site of the pain for 45 minutes. After "treatment," each patient was again asked to rate the level of pain from 0 to 10.

In all, 29 patients were given active magnets and 21 patients received inactive magnets. All but one of the patients rated their initial pain as an 8, 9, or 10. So scientists decided to focus on patients' final pain ratings. Here they are, grouped by the type of magnet used:<sup>1</sup>

---

Active: 0, 4, 7, 0, 4, 2, 5, 5, 3, 2, 2, 2, 3, 4, 6, 4, 3, 0, 2, 0, 4, 4, 5, 9, 10, 10, 10, 10, 7  
Inactive: 4, 7, 5, 8, 8, 6, 8, 10, 10, 6, 10, 8, 10, 10, 10, 10, 9, 9, 10, 10, 9

---

**What do the data tell us about whether the active magnets helped reduce pain? By the end of the chapter, you'll be ready to interpret the results of this study.**

## Introduction

You can hardly go a day without hearing the results of a statistical study. Here are some examples:

- The National Highway Traffic Safety Administration (NHTSA) reports that seat belt use in passenger vehicles increased from 83% in 2008 to 84% in 2009.<sup>2</sup>



- According to a recent survey, U.S. teens aged 13 to 18 spend an average of 26.8 hours per week online. Although 59% of the teens said that posting personal information or photos online is unsafe, 62% said they had posted photos of themselves.<sup>3</sup>

- A recent study suggests that lack of sleep increases the risk of catching a cold.<sup>4</sup>

- For their final project, two AP Statistics students showed that listening to music while studying decreased subjects' performance on a memory task.<sup>5</sup>

Can we trust these results? As you'll learn in this chapter, that depends on how the data were produced. Let's take a closer look at where the data came from in each of these studies.

Each year, the NHTSA conducts an *observational study* of seat belt use in vehicles. The NHTSA sends trained observers to record the actual behavior of people in vehicles at randomly selected locations across the country. The idea of an observational study is simple: you can learn a lot just by watching. Or by asking a few questions, as in the survey of teens' online habits. Harris Interactive conducted this survey using a "representative sample" of 655 U.S. 13- to 18-year-olds. Both of these studies use information from a *sample* to draw conclusions about some larger *population*. Section 4.1 examines the issues involved in sampling and surveys.

In the sleep and catching a cold study, 153 volunteers took part. They answered questions about their sleep habits over a two-week period. Then, researchers gave them a virus and waited to see who developed a cold. This was a complicated observational study. Compare this with the *experiment* performed by the AP Statistics students. They recruited 30 students and divided them into two groups of 15 by drawing names from a hat. Students in one group tried to memorize a list of words while listening to music. Students in the other group tried to memorize the same list of words while sitting in silence. Section 4.2 focuses on designing experiments.

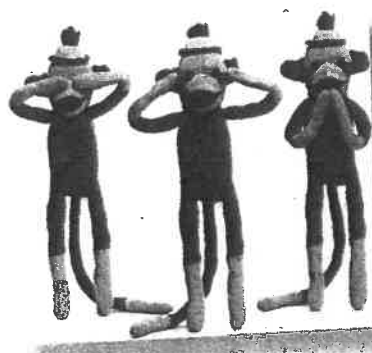
The goal of many statistical studies is to show that changes in one variable *cause* changes in another variable. In Section 4.3, we'll look at why establishing causation is so difficult, especially in observational studies. We'll also consider some of the ethical issues involved in planning and conducting a study.

Here's an Activity that gives you a preview of what lies ahead.

### • ACTIVITY *See no evil, hear no evil?*

- **MATERIALS:** Two index cards, each with 10 pairs of digits written on it (prepared by your teacher);
- clock, watch, or stopwatch to measure 30 seconds; and a coin for each pair of students

Confucius said, "I hear and I forget, I see and I remember, I do and I understand." Do people really remember what they see better than what they hear?<sup>6</sup> In this Activity, you will perform an experiment to try to find out.



1. Divide the class into pairs of students by drawing names from a hat.
2. Your teacher will give each pair two index cards with 10 sets of numbers on them. *Do not look at the numbers* until it is time for you to do the experiment.
3. Flip a coin to decide which of you is Student A and which is Student B. Shuffle the index cards and deal one face down to each partner.
4. Student A will attempt a memory task while Student B keeps time.

**Directions:** Study the pairs of numbers on the index card for 30 seconds. Then turn the card over. Recite the alphabet aloud (A, B, C, etc.). Then tell your partner (Student B) what you think the numbers on the card are. Student B will record how many pairs of numbers you recalled correctly.

5. Now it's Student B's turn to do a memory task while Student A records the data.

**Directions:** Your partner will read the pairs of numbers on your index card aloud three times slowly. Next, you will recite the alphabet aloud (A, B, C, etc.) and then tell your partner what you think the numbers on the card are. Student A will record how many pairs of numbers you recalled correctly.

6. Your teacher will scale and label axes on the board for parallel dotplots of the results. Plot the number of pairs you remembered correctly on the appropriate graph.
7. Did students in your class remember numbers better when they saw them or when they heard them? Give appropriate evidence to support your answer.
8. Based on the results of this experiment, can we conclude that people in general remember better when they see than when they hear? Why or why not?

## 4.1

# Sampling and Surveys

In Section 4.1,  
you'll learn about:

- The idea of a sample survey
- How to sample badly
- How to sample well: Random sampling
- Other sampling methods
- Inference for sampling
- Sample surveys: What can go wrong?

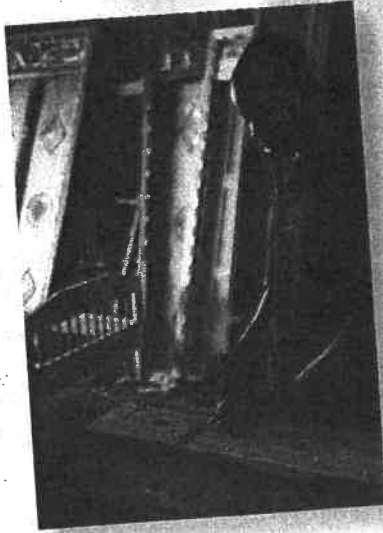
Suppose we want to find out what percent of young drivers in the United States text while driving. To answer the question, we will survey 16- to 20-year-olds who live in the United States and drive. Ideally, we would ask them all (take a *census*). But contacting every driver in this age group wouldn't be practical: it would take too much time and cost too much money. Instead, we put the question to a **sample** chosen to represent the entire **population** of young drivers.

### DEFINITION: Population and sample

The **population** in a statistical study is the entire group of individuals about which we want information.

A **sample** is the part of the population from which we actually collect information. We use information from a sample to draw conclusions about the entire population.

The distinction between population and sample is basic to statistics. To make sense of any sample result, you must know what population the sample represents. Here's an example that illustrates this distinction and also introduces some major uses of sampling.

**EXAMPLE****Sampling Hardwood and Humans**  
Populations and samples

**PROBLEM:** Identify the population and the sample in each of the following settings.

- (a) A furniture maker buys hardwood in large batches. The supplier is supposed to dry the wood before shipping (wood that isn't dry won't hold its size and shape). The furniture maker chooses five pieces of wood from each batch and tests their moisture content. If any piece exceeds 12% moisture content, the entire batch is sent back.
- (b) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.

**SOLUTION:**

- (a) The population is all the pieces of hardwood in a batch. The sample is the five pieces of wood that are selected from that batch and tested for moisture content.
- (b) Gallup's population is all adult U.S. residents. Their sample is the 1500 adults who actually respond to the survey questions.

**For Practice Try Exercise 1**

**The Idea of a Sample Survey**

We often draw conclusions about a whole population on the basis of a sample. Have you ever tasted a sample of ice cream and ordered a cone if the sample tastes good? Since ice cream is fairly uniform, the single taste represents the whole. Choosing a representative sample from a large and varied population (like all young U.S. drivers) is not so easy. The first step in planning a **sample survey** is to say exactly *what population* we want to describe. The second step is to say exactly *what we want to measure*, that is, to give exact definitions of our variables.

We reserve the term "sample survey" for studies that use an organized plan to choose a sample that represents some specific population, like the pieces of hardwood and the U.S. adults in the previous example. By our definition, the population in a sample survey can consist of people, animals, or things. Some people use the terms "survey" or "sample survey" to refer only to studies in which people are asked one or more questions, like the Gallup Poll of the last example. We'll avoid this more restrictive terminology.



Sample survey

**EXAMPLE****How Does the Current Population Survey Work?****A sample survey**

One of the most important government sample surveys in the United States is the monthly Current Population Survey (CPS). The CPS contacts about 60,000 households each month. It produces the monthly unemployment rate and much



other economic and social information. To measure unemployment, we must first specify the population we want to describe. The CPS defines its population as all U.S. residents (legal or not) 16 years of age and over who are civilians and are not in an institution such as a prison. The unemployment rate announced in the news refers to this specific population.

What does it mean to be “unemployed”? Someone who is not looking for work—for example, a full-time student—should not be called unemployed just because she is not working for pay. If you are chosen for the CPS sample, the interviewer first asks whether you are available to work and whether you actually looked for work in the past four weeks. If not, you are neither employed nor unemployed—you are not in the labor force.

If you are in the labor force, the interviewer goes on to ask about employment. If you did any work for pay or in your own business during the week of the survey, you are employed. If you worked at least 15 hours in a family business without pay, you are employed. You are also employed if you have a job but didn't work because of vacation, being on strike, or other good reason. An unemployment rate of 9.7% means that 9.7% of the sample was unemployed, using the exact CPS definitions of both “labor force” and “unemployed.”

The final step in planning a sample survey is to decide how to choose a sample from the population. Let's take a closer look at some good and not-so-good sampling methods.



## How to Sample Badly

How can we choose a sample that we can trust to represent the population? The easiest—but not the best—sampling method just chooses individuals who are close by. Suppose we're interested in finding out how long students at a large high school spent on homework last week, for example. We might go to the school library and ask the first 30 students we see about their homework time. A sample selected by taking the members of the population that are easiest to reach is called a **convenience sample**. Convenience samples often produce unrepresentative data.

### DEFINITION: Convenience sample

Choosing individuals who are easiest to reach results in a **convenience sample**.

What's wrong with our convenience sample of students in the library? It's unlikely that this sample represents the homework habits of all students at the school well. In fact, we'd expect the sample to overestimate the average homework time in the population since students who hang out in the library might tend to be more studious. This is **bias**: using a method that will consistently overestimate or underestimate the value you want to know.

**DEFINITION: Bias**

The design of a statistical study shows **bias** if it systematically favors certain outcomes.

**AP EXAM TIP** If you're asked to describe how the design of a study leads to bias, you're expected to identify the *direction* of the bias. Suppose you were asked, "Explain how using a convenience sample of students in your statistics class to estimate the proportion of all high school students who own a graphing calculator could result in bias." You might respond, "This sample would probably include a much higher proportion of students with a graphing calculator than in the population at large because a graphing calculator is required for the statistics class. That is, this method would probably lead to an overestimate of the actual population proportion."

Convenience samples are almost guaranteed to show bias. So are **voluntary response samples**, in which people decide whether to join the sample in response to an open invitation. Call-in, write-in, and many Internet polls rely on voluntary response samples. Unfortunately, people who choose to participate are usually not representative of any clearly defined population. Voluntary response samples attract people who feel strongly about the issue in question, often in the same direction. That leads to bias.

The Internet brings voluntary response samples to the computer nearest you. Visit [www.misterpoll.com](http://www.misterpoll.com) to become part of the sample in any of dozens of online polls. As the site says, "None of these polls are 'scientific,' but do represent the collective opinion of everyone who participates." Unfortunately, such polls don't tell you anything about the views of the population at large.

**DEFINITION: Voluntary response sample**

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples show bias because people with strong opinions (often in the same direction) are most likely to respond.

Write-in and call-in opinion polls are almost sure to lead to strong bias. In fact, only about 15% of the public has ever responded to a call-in poll, and these tend to be the same people who call radio talk shows. That's not a representative sample of the population as a whole.

**EXAMPLE**

### *Illegal Immigrants and Driver's Licenses*

Online polls

Former CNN commentator Lou Dobbs doesn't like illegal immigration. One of his shows was largely devoted to attacking a proposal to offer driver's licenses to illegal immigrants. During the show, Mr. Dobbs invited his viewers to go to [loudobbs.com](http://loudobbs.com) to vote on the question "Would you be more or less likely to vote for a presidential candidate who supports giving drivers' licenses to illegal aliens? The result: 97% of the 7350 people who voted by the end of the show said "Less likely."

**PROBLEM:** What type of sample did Mr. Dobbs use in his poll? Explain how this sampling method could lead to bias in the poll results.

**SOLUTION:** Mr. Dobbs used a voluntary response sample: people chose to go online and respond. Those who voted were viewers of Mr. Dobbs's program, which means that they are likely to support his views. The 97% poll result is probably an extreme overestimate of the percent of people in the population who would be less likely to support a presidential candidate with this position.

For Practice Try Exercise 5



### CHECK YOUR UNDERSTANDING

For each of the following situations, identify the sampling method used. Then explain how the sampling method could lead to bias.

1. A farmer brings a juice company several crates of oranges each week. A company inspector looks at 10 oranges from the top of each crate before deciding whether to buy all the oranges.
2. The ABC program *Nightline* once asked whether the United Nations should continue to have its headquarters in the United States. Viewers were invited to call one telephone number to respond "Yes" and another for "No." There was a charge for calling either number. More than 186,000 callers responded, and 67% said "No."

## How to Sample Well: Random Sampling

In a voluntary response sample, people choose whether to respond. In a convenience sample, the interviewer makes the choice. In both cases, personal choice produces bias. The statistician's remedy is to allow impersonal chance to choose the sample. A sample chosen by chance rules out both favoritism by the sampler and self-selection by respondents. **Random sampling**, the use of chance to select a sample, is the central principle of statistical sampling.

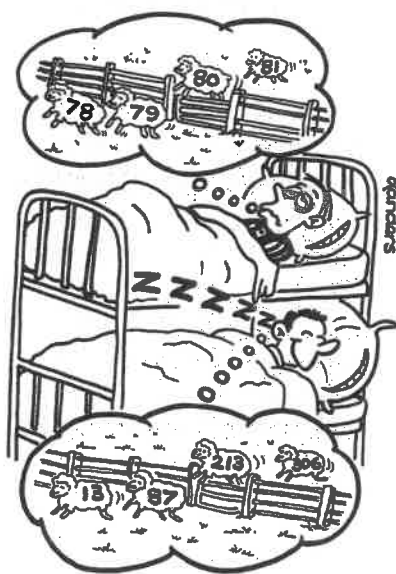
The simplest way to use chance to select a sample is to place names in a hat (the population) and draw out a handful (the sample). This is the idea of a **simple random sample**.

### Random sampling

In everyday life, some people use the word "random" to mean haphazard, as in, "That's so random." In statistics, random means "due to chance." Don't say that a sample was chosen at random if a chance process wasn't used to select the individuals.

#### DEFINITION: Simple random sample

A **simple random sample (SRS)** of size  $n$  consists of  $n$  individuals from the population chosen in such a way that every set of  $n$  individuals has an equal chance to be the sample actually selected.



Statisticians fall asleep faster by taking a random sample of sheep.

An SRS not only gives each individual an equal chance to be chosen but also gives every possible sample an equal chance to be chosen. There are other random sampling methods that give each individual, but not each sample, an equal chance. Exercise 26 describes one such method. For now, let's think about how to actually select an SRS.

When you think of an SRS, picture drawing names from a hat to remind yourself that an SRS doesn't favor any part of the population. That's why an SRS is a better method of choosing samples than convenience or voluntary response samples. But writing names on slips of paper and drawing them from a hat doesn't work as well if the population is large. Think about the Current Population Survey, for example,

which must draw a sample of size 60,000 every month from the population of over 117 million U.S. households.

In practice, people use random numbers generated by a computer or calculator to choose samples. If you don't have technology handy, you can use a **table of random digits**.

**DEFINITION: Table of random digits**

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these properties:

- Each entry in the table is equally likely to be any of the 10 digits 0 through 9.
- The entries are independent of each other. That is, knowledge of one part of the table gives no information about any other part.

Table D at the back of the book is a table of random digits. It begins with the digits 19223950340575628713. To make the table easier to read, the digits appear in groups of five and in numbered rows. The groups and rows have no meaning—the table is just a long list of randomly chosen digits. There are two steps in using the table to choose a simple random sample.

**How to Choose an SRS Using Table D**

**STEP 1: LABEL.** Give each member of the population a numerical label of the *same length*.

**STEP 2: TABLE.** Read consecutive groups of digits of the appropriate length from Table D.

Your sample contains the individuals whose labels you find.

Always use the shortest labels that will cover your population. For instance, you can label up to 100 individuals with two digits: 01, 02, . . . , 99, 00. As standard practice, we recommend that you begin with label 1 (or 01 or 001, as needed). Reading groups of digits from the table gives all individuals the same chance to be chosen because all labels of the same length have the same chance to be found in the table. For example, any pair of digits in the table is equally likely to be any of the 100 possible labels 01, 02, . . . , 99, 00. Ignore any group of digits that was not used as a label or that duplicates a label already in the sample. You can read digits from Table D in any order—across a row, down a column, and so on—because the table has no order. We recommend reading across rows from left to right.



**EXAMPLE***Spring Break!*

## Choosing an SRS with Table D



The school newspaper is planning an article on family-friendly places to stay over spring break at a nearby beach town. The editors intend to call 4 randomly chosen hotels to ask about their amenities for families with children. They have an alphabetized list of all 28 hotels in the town.

**PROBLEM:** Use Table D at line 130 to choose an SRS of 4 hotels for the editors to call.

**SOLUTION:** We'll use the two-step process for selecting an SRS using Table D.

**Step 1: Label.** Two digits are needed to label the 28 resorts. We have added labels 01 to 28 to the alphabetized list of resorts below.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 Banyan Tree	11 Diplomat	18 Sandpiper	25 Tradewinds
05 Beach Castle	12 Holiday Inn	19 Sea Castle	26 Tropical Breeze
06 Best Western	13 Lime Tree	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda

**Step 2: Table.** To use Table D, start at the left-hand side of line 130 and read two-digit groups. Skip any groups that aren't between 01 and 28, as well as any repeated groups. Continue until you have chosen four resorts. Here is the beginning of line 130:

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups are

69	05	16	48	17	87	17	40	95	17
Skip	✓	✓	Skip	✓	Skip	Skip	Skip	Skip	Skip
Too big			Too big		Too big	Repeat	Too big	Too big	Repeat

We skip 5 of these 10 groups because they are too high (over 28) and 2 because they are repeats (both 17s). The hotels labeled 05, 16, and 17 go into the sample. We need one more hotel to complete the sample. The remaining 10 two-digit groups in line 130 are

84	53	40	64	89	87	20	19	72	45
Skip	Skip	Skip	Skip	Skip	Skip	✓			
Too big									

Our SRS of 4 hotels for the editors to contact is: 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.

**For Practice Try Exercise 11**

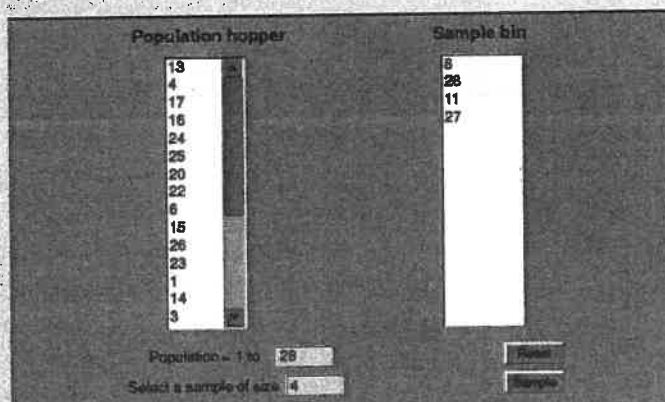
As you saw in the previous example, using Table D to select an SRS can be time-consuming. The *Simple Random Sample* applet can help you quickly choose an SRS for populations of up to 500 individuals. Your calculator can do even better.



## TECHNOLOGY CORNER Choosing an SRS

You can use the *Simple Random Sample* applet at [www.whfreeman.com/tps4e](http://www.whfreeman.com/tps4e) or your graphing calculator to select the SRS of 4 hotels in the previous example.

**Simple Random Sample applet:** Enter 28 in the “Population = 1 to” box and 4 in the “Select a sample of size” box, click “Reset,” and click “Sample.” Figure 4.1 shows the result of one sample.



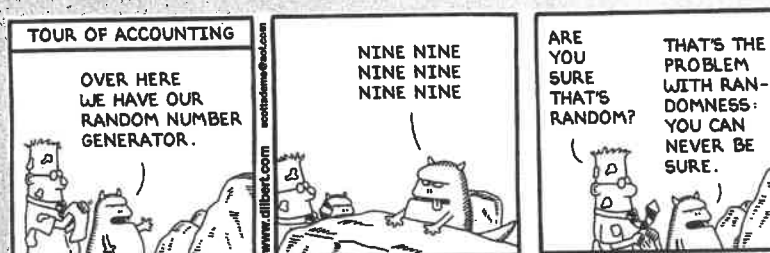
**FIGURE 4.1** The *Simple Random Sample* applet used to choose an SRS of size  $n = 4$  from a population of size 28.

**Graphing calculator:**

1. Check that your calculator's random number generator is working properly.

TI-83/84: Press **MATH**, then select PRB and 5:randInt(. Complete the command randInt(1,28) and press **ENTER**.

TI-89: Press **CATALOG**, then **F3** (Flash Apps), and choose **randInt** (. Complete the command **T1Stat:randInt(1,28)** and press **ENTER**.



Compare results with your classmates. If several students got the same number, you'll need to seed your calculator's random number generator with different numbers before you proceed. Directions for doing this are given in the *Annotated Teachers Edition*.

2. Now do `randInt(1,28)` again. Keep pressing **ENTER** until you have chosen four different labels. (Notice the repeated label "25" in the TI-84 screen shot.)

```
randInt(1,28)
```

F1=	F2=	F3=	F4=	F5=	F6=
Tool	Refers	Calc	Other	PrsView	Chart, BP
<pre> ■ tistat.randint(1,28) 19 ■ tistat.randint(1,28) 25 ■ tistat.randint(1,28) 26 ■ tistat.randint(1,28) 11 ■ tistat.randint(1,28)           </pre>					
MAIN	END AUTO	FINCH	4/25		

TI-Nspire instructions in Appendix B

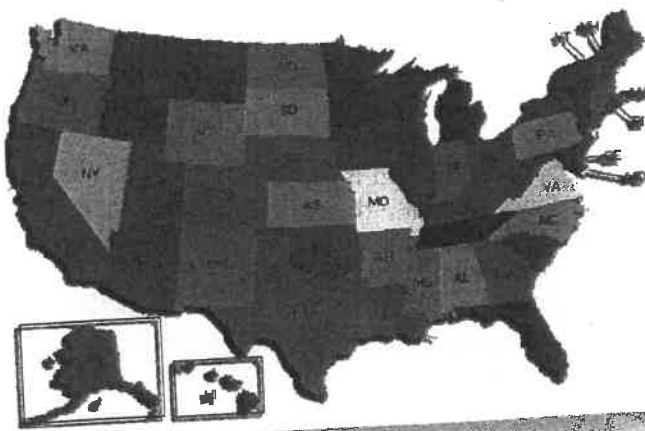
To see how software speeds up choosing an SRS, try the random number generator at [www.random.org](http://www.random.org). The Research Randomizer at [www.randomizer.org](http://www.randomizer.org) is another option. Click on "Randomize" and fill in the boxes. You can even ask the Randomizer to arrange your sample in order.

There's an important difference between the samples produced by the *Simple Random Sample* applet and the calculator's `randInt` command in the Technology Corner. The applet sampled *without replacement* from the "Population hopper," but the calculator sampled *with replacement* from the specified population. As a result, the calculator sometimes selects the same number more than once in a given sample. To deal with this problem, you can generate additional random numbers as needed to replace any repeats. Or you can use a method other than `randInt` (like `randSamp` on the TI-Nspire) to sample without replacement. Refer to your device's reference manual or ask your teacher.

We can trust results from an SRS, as well as from other types of random samples that we will meet later, because the use of impersonal chance avoids bias. Online and call-in polls also produce samples. But we can't trust results from these samples because they are chosen in ways that invite bias. *The first question to ask about any sample is whether it was chosen at random.*

### ACTIVITY *How large is a typical U.S. state?*

In this Activity, you and your classmates will use two different sampling methods to estimate the average area (in square miles) of the 50 states.



1. Use the map shown to choose a sample of 5 states to estimate the average (mean) land area. You have 15 seconds.
2. Refer to the table of land areas on page N/DS-5. Find the mean land area for your sample.
3. Make a class dotplot of the mean land areas from Step 2.
4. Use a line of Table D assigned by your teacher to choose an SRS of 5 states from the table on page N/DS-5. Find the mean land area for this sample.
5. Make a class dotplot of the mean land areas from Step 4 right above your dotplot from Step 3.

6. How do the class's estimates using the two methods compare? What advantage(s) does random sampling provide?

## Other Sampling Methods

The basic idea of sampling is straightforward: take an SRS from the population and use your sample results to gain information about the population. Unfortunately, it's usually very difficult to actually get an SRS from the population of interest. Imagine trying to get a simple random sample of all the batteries produced in one day at a factory. Or an SRS of all U.S. high school students. In either case, it's just not practical to choose an SRS. For starters, it would be difficult to obtain an accurate list of the population from which to draw the sample. It would also be very time-consuming to collect data from each individual that's randomly selected. Sometimes, there are also statistical advantages to using more complex sampling methods.

One of the most common alternatives to an SRS involves sampling important groups (called **strata**) within the population separately. Then these separate “sub-samples” are combined to form one **stratified random sample**. This method works best when the individuals within each stratum are similar to one another in a way that affects the variable being measured but there are large differences between strata.

**DEFINITION: Stratified random sample and strata**

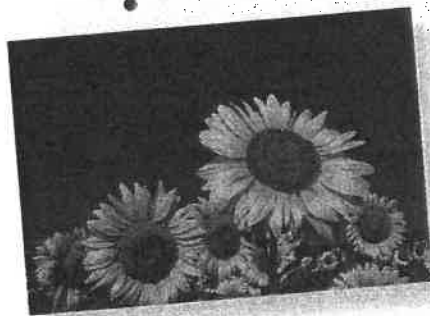
To select a **stratified random sample**, first classify the population into groups of similar individuals, called **strata**. Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

Unfortunately, “stratified random sample” has the same initials as “simple random sample.” Only a simple random sample gets abbreviated SRS, however.

Choose the strata based on facts known before the sample is taken. For example, in a study of sleep habits on school nights, the population of students in a large high school might be divided into freshman, sophomore, junior, and senior strata. In a preelection poll, a population of election districts might be divided into urban, suburban, and rural strata. *If the individuals in each stratum are less varied than the population as a whole, a stratified random sample can produce better information about the population than an SRS of the same size.* Not convinced? Try the following Activity.

### • ACTIVITY *Sampling sunflowers*

- A British farmer grows sunflowers for making sunflower oil. Her field is arranged in a grid pattern, with 10 rows and 10 columns as shown in the figure. Irrigation ditches run along the top and bottom of the field, as shown. The farmer would like to estimate the number of healthy plants in the field so she can project how much money she'll make from selling them. It would take too much time to count the plants in all 100 squares, so she'll accept an estimate based on a sample of 10 squares.



1. Use Table D or technology to take a simple random sample of 10 grid squares. Record the location (for example, B6) of each square you select.

2. This time, you'll take a stratified random sample using the rows as strata. Use Table D or technology to randomly select one square from each (horizontal) row. Record the location of each square.

3. Now, take a stratified random sample using the *columns* as strata. Use Table D or technology to randomly select one square from each (vertical) column. Record the location of each square.

4. The table on page N/DS-6 gives the actual number of sunflowers in each grid square. Use the information provided to calculate your estimate of the mean number of sunflowers per square for each of your samples in Steps 1, 2, and 3.

[illegible]



5. Make comparative dotplots showing the mean number of sunflowers obtained using the three different sampling methods for all members of the class. Describe any similarities and differences you see.
6. Your teacher will provide you with the mean number of sunflowers in the population of all 100 grid squares in the field. How did the three sampling methods do?

The following example describes an interesting use of stratified random samples in the music business.

## EXAMPLE

### *Who Wrote That Song?*

#### Using a stratified random sample

A radio station that broadcasts a piece of music owes a royalty to the composer. The organization of composers (called ASCAP) collects these royalties for all its members by charging stations a license fee for the right to play members' songs. ASCAP has four million songs in its catalog and collects \$435 million in fees each year. How should ASCAP distribute this income among its members? By sampling: ASCAP tapes about 60,000 hours from the 53 million hours of local radio programs across the country each year.

Radio stations are stratified by type of community (metropolitan, rural), geographic location (New England, Pacific, etc.), and the size of the license fee paid to ASCAP, which reflects the size of the audience. In all, there are 432 strata. Tapes are made at random hours of randomly selected members of each stratum. The tapes are reviewed by experts who can recognize almost every piece of music ever written, and the composers are then paid according to their popularity.<sup>7</sup>

Although a stratified random sample can sometimes give more precise information about a population than an SRS, both sampling methods are hard to use when populations are large and spread out over a wide area. In that situation, we'd prefer a method that selects groups of individuals that are "near" one another. That's the idea of a **cluster sample**.

#### **DEFINITION: Cluster sample and clusters**

To take a **cluster sample**, first divide the population into smaller groups. Ideally, these **clusters** should mirror the characteristics of the population. Then choose an SRS of the clusters. All individuals in the chosen clusters are included in the sample.

In a cluster sample, some people take an SRS from each cluster rather than including all members of the cluster.

Imagine a large high school that assigns its students to homerooms alphabetically by last name. The school administration is considering a new schedule and would like student input. Administrators decide to survey 200 randomly selected students. It would be hard to track down an SRS of 200 students, so the administration opts for a cluster sample of homerooms. The principal (who knows some

You might say that strata are ideally “similar within, but different between,” while clusters are ideally “different within, but similar between.”

statistics) takes a simple random sample of 8 homerooms and gives the survey to all 25 students in each homeroom.

Cluster samples are often used for practical reasons, as in the school survey example. They don’t offer the statistical advantage of better information about the population that stratified random samples do. That’s because clusters are often chosen for ease or convenience, so they may have as much variability as the population itself. Be sure you understand the difference between strata and clusters. We want each stratum to contain similar individuals, and for there to be large differences between strata. For a cluster sample, we’d *like* each cluster to look just like the population, but on a smaller scale.

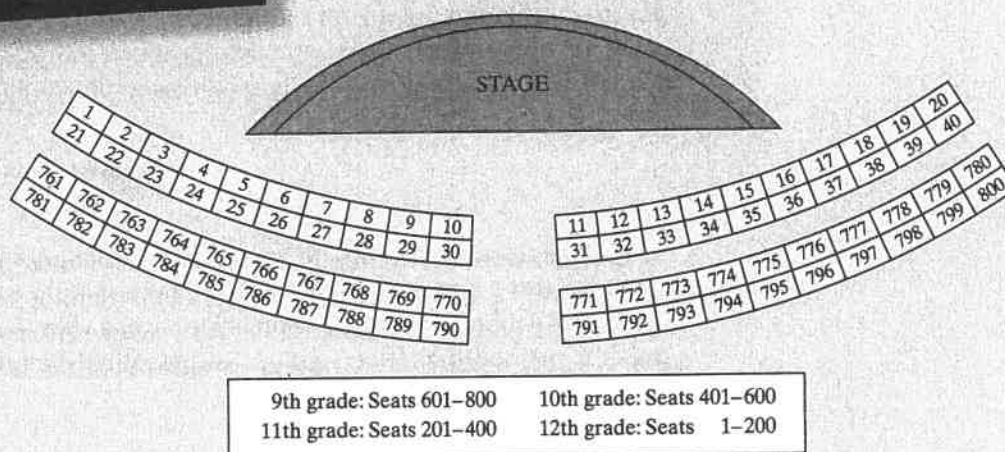
## EXAMPLE

### *Sampling at a School Assembly* Strata or clusters?



The student council wants to conduct a survey during the first five minutes of an all-school assembly in the auditorium about use of the school library. They would like to announce the results of the survey at the end of the assembly. The student council president asks your statistics class to help carry out the survey.

**PROBLEM:** There are 800 students present at the assembly. A map of the auditorium is shown below. Note that students are seated by grade level and that the seats are numbered from 1 to 800.



Describe how you would use each of the following sampling methods to select 80 students to complete the survey.

- Simple random sample
- Stratified random sample
- Cluster sample

**SOLUTION:**

- To take an SRS, we need to choose 80 of the seat numbers at random. Use `randInt(1,800)` on your calculator until 80 different seats are selected. Then give the survey to the students in those seats.

Can you give an advantage and a disadvantage of each sampling method used in the example?

(b) The students in the assembly are seated by grade level. Since students' library use might be similar within grade levels but different across grade levels, we'll use the grade level seating areas as our strata. Within each grade's seating area, we'll select 20 seats at random. For the 9th grade, use  $\text{randInt}(601, 800)$  to select 20 seats. Use  $\text{randInt}(401, 600)$  to pick 20 sophomore seats,  $\text{randInt}(201, 400)$  to get 20 junior seats, and  $\text{randInt}(1, 200)$  to choose 20 senior seats. Give the survey to the students in the selected seats.

(c) With the way students are seated, each column of seats from the stage to the back of the auditorium could be used as a cluster. Note that each cluster contains students from all four grade levels, so each should represent the population well. Since there are 20 clusters, each with 40 seats, we need to choose 2 clusters at random to get 80 students for the survey. Use  $\text{randInt}(1, 20)$  to select two clusters, and then give the survey to all 40 students in each column of seats.

### For Practice Try Exercise 21

Most large-scale sample surveys use *multistage samples* that combine two or more sampling methods. For example, the Current Population Survey's monthly survey of 60,000 households uses the following design:

**Stage 1:** Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. Select a sample of 756 PSUs. This sample includes the 428 PSUs with the largest population and a stratified random sample of 328 of the others.

**Stage 2:** Divide each PSU selected into smaller areas called "neighborhoods." Stratify the neighborhoods using ethnic and other information, and take a stratified random sample of the neighborhoods in each PSU.

**Stage 3:** Sort the housing units in each neighborhood into clusters of four nearby units. Interview the households in a random sample of these clusters.

Analyzing data from sampling methods more complex than an SRS takes us beyond basic statistics. But the SRS is the building block of more elaborate methods, and the principles of analysis remain much the same for these other methods.

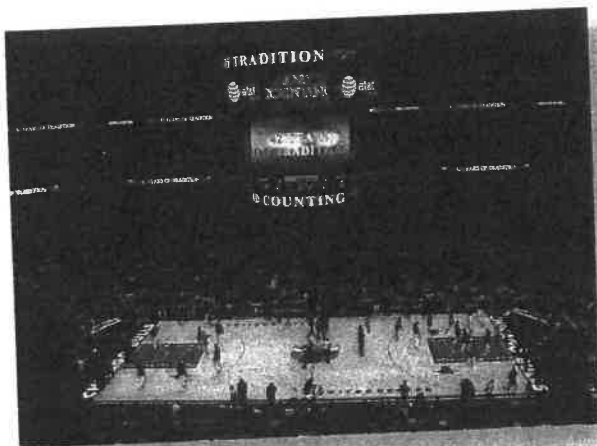


### CHECK YOUR UNDERSTANDING

The manager of a sports arena wants to learn more about the financial status of the people who are attending an NBA basketball game. He would like to give a survey to a representative sample of the more than 20,000 fans in attendance. Ticket prices for the game vary a

great deal: seats near the court cost over \$100 each, while seats in the top rows of the arena cost \$25 each. The arena is divided into 30 numbered sections, from 101 to 130. Each section has rows of seats labeled with letters from A (nearest the court) to ZZ (top row of the arena).

1. Explain why it might be difficult to give the survey to an SRS of 200 fans.
2. Which would be a better way to take a stratified random sample of fans: using the lettered rows or the numbered sections as strata? Explain.
3. Which would be a better way to take a cluster sample of fans: using the lettered rows or the numbered sections as clusters? Explain.



## Inference for Sampling

### Inference

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called **inference** because we *infer* information about the population from what we *know* about the sample. Inference from convenience samples or voluntary response samples would be misleading because these methods of choosing a sample are biased. We are almost certain that the sample does *not* fairly represent the population. *The first reason to rely on random sampling is to eliminate bias in selecting samples from the list of available individuals.*

Still, it is unlikely that results from a random sample are exactly the same as for the entire population. Sample results, like the unemployment rate obtained from the monthly Current Population Survey, are only estimates of the truth about the population. If we select two samples at random from the same population, we will almost certainly choose different individuals. So the sample results will differ somewhat, just by chance. Properly designed samples avoid systematic bias. But their results are rarely exactly correct, and we expect them to vary from sample to sample.

### • ACTIVITY *Results may vary . . .*

- **MATERIALS:** At least 200 colored chips with exactly 60% of a particular color (say, red); large bag or other container that students can't see through
- 

In this Activity, you will explore how the results of repeated random samples vary in relation to the population truth.

1. Your teacher will prepare a large population of colored chips with 60% of a particular color.
2. One at a time, each student in the class will take an SRS of 20 chips and record the proportion of chips obtained that are red. The chips should be returned to the container before the next student chooses a sample.
3. Make a class dotplot of the sample proportions for this color of chip. Where is the graph centered? How much do the sample proportions vary around the center? Is there a clear shape?

### Margin of error

Why can we trust random samples? As the previous Activity illustrates, the results of random sampling don't change haphazardly from sample to sample. Because we deliberately use chance, the results obey the laws of probability that govern chance behavior. These laws allow us to say how likely it is that sample results are close to the truth about the population. *The second reason to use random sampling is that the laws of probability allow trustworthy inference about the population.* Results from random samples come with a **margin of error** that sets bounds on the size of the likely error. We will discuss the details of inference for sampling later.

One point is worth making now: *larger random samples give better information about the population than smaller samples.* For instance, in the Activity, you would get a better estimate of the proportion of red chips in the popula-



tion using random samples of 40 chips than 20 chips. By taking a very large sample, you can be confident that the sample result is very close to the truth about the population.

The Current Population Survey contacts about 60,000 households, so we'd expect its estimate of the national unemployment rate to be within about 0.1% of the actual population value. Opinion polls that contact 1000 or 1500 people give less precise results—we expect the sample result to be within about 3% of the actual population percent with a given opinion. Of course, only samples chosen by chance carry this guarantee. Lou Dobbs's online sample tells us little about overall American public opinion even though 7350 people clicked a response.

## Sample Surveys: What Can Go Wrong?

Random sampling eliminates bias in choosing a sample. But even a large random sample will give a result that differs from the truth about the population. This “sampling variability” is described by the margin of error that comes with most poll results. So once we see the magic words *randomly selected*, do we know we have trustworthy information? It certainly beats voluntary response, but not always by as much as we might hope. Sampling in the real world is more complex and less reliable than choosing an SRS from a list of names in a textbook exercise.

Most sample surveys are affected by errors in addition to sampling variability. These errors can introduce bias that makes a survey result meaningless. Good sampling technique includes the art of reducing all sources of error. Let's look at the two main sources of errors in sample surveys: sampling errors and nonsampling errors.

**Sampling errors** The margin of error tells us how much sampling variability to expect, and we can control it by choosing the size of our random sample. It doesn't tell us about sampling errors, mistakes made in the process of taking a sample that could lead to inaccurate information about the population. One source of sampling error is the use of *bad sampling methods*, such as voluntary response. We can avoid bad methods. Other sampling errors are not so easy to handle. Sampling often begins with a list of individuals from which we will draw our sample. This list is called the **sampling frame**. Ideally, the sampling frame should list every individual in the population. Because a list of the entire population is rarely available, most samples suffer from some degree of **undercoverage**.

Sampling frame

### DEFINITION: Undercoverage

**Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample.

A sample survey of households, for example, will miss not only homeless people but prison inmates and students in dormitories. An opinion poll conducted by calling landline telephone numbers will miss households that have only cell phones as well as households without a phone. The results of national sample surveys therefore have some bias due to undercoverage if the people not covered differ systematically from the rest of the population.

Sampling errors in careful sample surveys are usually quite small. The real problems start when someone picks up (or doesn't pick up) the phone. Now non-sampling errors take over.

**Nonsampling errors** Nonsampling errors are those that can plague even a census. One of the most serious sources of bias in sample surveys is **nonresponse**, which occurs when a selected individual cannot be contacted or refuses to cooperate. Nonresponse to sample surveys often exceeds 50%, even with careful planning and several callbacks. Because nonresponse is higher in urban areas, most sample surveys substitute other people in the same area to avoid favoring rural areas in the final sample. If the people contacted differ from those who are rarely at home or who refuse to answer questions, some bias remains. For example, retired people may be more likely to respond, which would give their opinions more weight. In a poll about Social Security reform, this could give a misleading impression of the population's views.

#### DEFINITION: Nonresponse

**Nonresponse** occurs when an individual chosen for the sample can't be contacted or refuses to participate.

*Some students misuse the term “voluntary response” to explain why certain individuals don't respond in a sample survey. Their idea is that participation in the survey is optional (voluntary), so anyone can refuse to take part. What the students are actually describing is nonresponse. Think about it this way: nonresponse can occur only after a sample has been selected. In a voluntary response sample, every individual has opted to take part, so there won't be any nonresponse.*



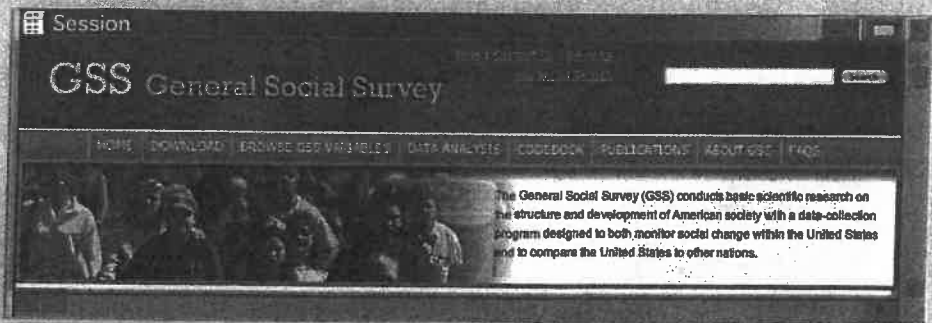
### EXAMPLE

#### *The ACS, GSS, and Opinion Polls*

##### **How bad is nonresponse?**

The Census Bureau's American Community Survey (ACS) has the lowest nonresponse rate of any poll we know: only about 1% of the households in the sample refuse to respond. The overall nonresponse rate, including “never at home” and other causes, is just 2.5%.<sup>8</sup> This monthly survey of about 250,000 households replaces the “long form” that in the past was sent to some households in the every-ten-years national census. Participation in the ACS is mandatory, and the Census Bureau follows up by telephone and then in person if a household doesn't return the mail questionnaire.

The University of Chicago's General Social Survey (GSS) is the nation's most important social science survey (see Figure 4.2). The GSS contacts its sample in person, and it is run by a university. Despite these advantages, its most recent survey had a 30% rate of nonresponse.

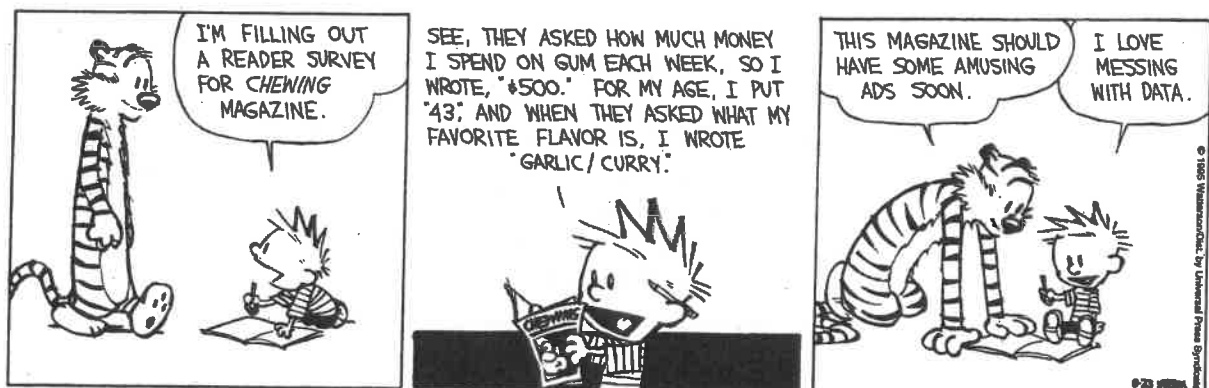


**FIGURE 4.2** The home page of the General Social Survey at the University of Chicago's National Opinion Research Center. The GSS has tracked opinions about a wide variety of issues since 1972.

What about opinion polls by news media and opinion-polling firms? We don't know their rates of nonresponse because they won't say. That's a bad sign. The Pew Research Center for the People and the Press imitated a careful random digit dialing survey and published the results: over 5 days, the survey reached 76% of the households in its chosen sample, but "because of busy schedules, skepticism and outright refusals, interviews were completed in just 38% of households that were reached." Combining households that could not be contacted with those who did not complete the interview gave a nonresponse rate of 73%.<sup>9</sup>

### Response bias

Another type of nonsampling error occurs when someone gives an incorrect response. A systematic pattern of incorrect responses in a sample survey leads to **response bias**. People know that they should take the trouble to vote, for example, so many who didn't vote in the last election will tell an interviewer that they did. The race or gender of the interviewer can influence responses to questions about race relations or attitudes toward feminism. Answers to questions that ask people to recall past events are often inaccurate because of faulty memory. "Have you visited a dentist in the last 6 months?" will often draw a "Yes" from someone who last visited a dentist 8 months ago.<sup>10</sup> Careful training of interviewers and careful supervision to avoid variation among the interviewers can reduce response bias. Good interviewing technique is another aspect of a well-done sample survey.



## Wording of questions

The **wording of questions** is the most important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and changes in wording can greatly change a survey's outcome. Even the order in which questions are asked matters.

**EXAMPLE***How Do Americans Feel about Illegal Immigrants?***Question wording matters**

"Should illegal immigrants be prosecuted and deported for being in the U.S. illegally, or shouldn't they?" Asked this question in an opinion poll, 69% favored deportation. But when the very same sample was asked whether illegal immigrants who have worked in the United States for two years "should be given a chance to keep their jobs and eventually apply for legal status," 62% said that they should. Different questions give quite different impressions of attitudes toward illegal immigrants.

*Don't trust the results of a sample survey until you have read the exact questions asked.* The amount of nonresponse and the date of the survey are also important. Good statistical design is a part, but only a part, of a trustworthy survey.

**THINK  
ABOUT  
IT**

**Does the order matter?** Ask a sample of college students these two questions:

- "How happy are you with your life in general?" (Answers on a scale of 1 to 5)  
 "How many dates did you have last month?"

There is almost no association between responses to the two questions when asked in this order. It appears that dating has little to do with happiness. Reverse the order of the questions, however, and a much stronger association appears: college students who say they had more dates tend to give higher ratings of happiness about life. Asking a question that brings dating to mind makes dating success a big factor in happiness.

**CHECK YOUR UNDERSTANDING**

1. Each of the following is a source of error in a sample survey. Label each as *sampling error* or *nonsampling error*, and explain your answers.

- The telephone directory is used as a sampling frame.
- The person cannot be contacted in five calls.
- Interviewers choose people walking by on the sidewalk to interview.

2. A survey paid for by makers of disposable diapers found that 84% of the sample opposed banning disposable diapers. Here is the actual question:

It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?<sup>11</sup>

Explain how the wording of the question could result in bias. Be sure to specify the direction of the bias.



## SECTION 4.1

## Summary

- A **sample survey** selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data from the sample. It is important to specify exactly what population you are interested in and what variables you will measure.
- **Random sampling** uses chance to select a sample.
- The basic random sampling method is a **simple random sample (SRS)**. An SRS gives every possible sample of a given size the same chance to be chosen. Choose an SRS by labeling the members of the population and using **random digits** to select the sample. Technology can automate this process.
- To choose a **stratified random sample**, divide the population into **strata**, groups of individuals that are similar in some way that might affect their responses. Then choose a separate SRS from each stratum.
- To choose a **cluster sample**, divide the population into groups, or **clusters**. Randomly select some of these clusters. All the individuals in the chosen clusters are included in the sample.
- Failure to use random sampling often results in **bias**, or systematic errors in the way the sample represents the population. **Voluntary response samples**, in which the respondents choose themselves, and **convenience samples**, in which individuals close by are included in the sample, are particularly prone to large bias.
- **Sampling errors** come from the act of choosing a sample. Random sampling error and **undercoverage** are common types of sampling error. Undercoverage occurs when some members of the population are left out of the **sampling frame**, the list from which the sample is actually chosen.
- The most serious errors in most careful surveys, however, are **nonsampling errors**. These have nothing to do with choosing a sample—they are present even in a census. The single biggest problem for sample surveys is **nonresponse**: people can't be contacted or refuse to answer. Incorrect answers by respondents can lead to **response bias**. Finally, the exact **wording of questions** has a big influence on the answers.

## 4.1 TECHNOLOGY CORNER

Choosing an SRS..... page 214

TI-Nspire instructions in Appendix B

## SECTION 4.1

## Exercises

pg 208

1. **Students as customers** A high school's student newspaper plans to survey local businesses about the importance of students as customers. From telephone book listings, the newspaper staff chooses 150 businesses at random. Of these, 73 return the questionnaire mailed by the staff. Identify the population and the sample.

2. **Student archaeologists** An archaeological dig turns up large numbers of pottery shards, broken stone tools, and other artifacts. Students working on the project classify each artifact and assign it a number. The counts in different categories are important for understanding the site, so the project director chooses 2% of the artifacts at random and checks the students' work. Identify the population and the sample.

3. **Sampling stuffed envelopes** A large retailer prepares its customers' monthly credit card bills using an automatic machine that folds the bills, stuffs them into envelopes, and seals the envelopes for mailing. Are the envelopes completely sealed? Inspectors choose 40 envelopes from the 1000 stuffed each hour for visual inspection. Identify the population and the sample.

4. **Customer satisfaction** A department store mails a customer satisfaction survey to people who make credit card purchases at the store. This month, 45,000 people made credit card purchases. Surveys are mailed to 1000 of these people, chosen at random, and 137 people return the survey form. Identify the population and the sample.

pg 210

5. **Call the shots** A newspaper advertisement for an upcoming TV show said: "Should handgun control be tougher? You call the shots in a special call-in poll tonight. If yes, call 1-900-720-6181. If no, call 1-900-720-6182. Charge is 50 cents for the first minute." Explain why this opinion poll is almost certainly biased.

6. **Explain it to the congresswoman** You are on the staff of a member of Congress who is considering a bill that would provide government-sponsored insurance for nursing-home care. You report that 1128 letters have been received on the issue, of which 871 oppose the legislation. "I'm surprised that most of my constituents oppose the bill. I thought it would be quite popular," says the congresswoman. Are you convinced that a majority of the voters oppose the bill? How would you explain the statistical issue to the congresswoman?

www

7. **Instant opinion** A recent online poll posed the question "Should female athletes be paid the same as men for the work they do?" In all, 13,147 (44%) said "Yes," 15,182 (50%) said "No," and the remaining 1448 said "Don't know." In spite of the large sample size for this survey, we can't trust the result. Why not?

8. **Online polls** In June 2008, *Parade* magazine posed the following question: "Should drivers be banned from using all cell phones?" Readers were encouraged to vote online at *parade.com*. The July 13, 2008, issue of *Parade* reported the results: 2407 (85%) said "Yes" and 410 (15%) said "No."

(a) What type of sample did the *Parade* survey obtain?

(b) Explain why this sampling method is biased. Is 85% probably higher or lower than the true percent of all adults who believe that cell phone use while driving should be banned? Why?

9. **Sleepless nights** How much sleep do high school students get on a typical school night? An interested student designed a survey to find out. To make data collection easier, the student surveyed the first 100 students to arrive at school on a particular morning. These students reported an average of 7.2 hours of sleep on the previous night.

(a) What type of sample did the student obtain?

(b) Explain why this sampling method is biased. Is 7.2 hours probably higher or lower than the true average amount of sleep last night for all students at the school? Why?

10. **Sampling at the mall** You have probably seen the mall interviewer, approaching people passing by with clipboard in hand. Explain why even a large sample of mall shoppers would not provide a trustworthy estimate of the current unemployment rate.

pg 213

11. **Do you trust the Internet?** You want to ask a sample of high school students the question "How much do you trust information about health that you find on the Internet—a great deal, somewhat, not much, or not at all?" You try out this and other questions on a pilot group of 5 students chosen from your class. The class members are listed at top right.

(a) Explain how you would use a line of Table D to choose an SRS of 5 students from the following list. Explain your method clearly enough for a classmate to obtain your results.

(b) Use line 107 to select the sample. Show how you use each of the digits.

Anderson	Deng	Glaus	Nguyen	Samuels
Arroyo	De Ramos	Helling	Palmiero	Shen
Batista	Drasin	Husain	Percival	Tse
Bell	Eckstein	Johnson	Prince	Velasco
Burke	Fernandez	Kim	Puri	Wallace
Cabrera	Fullmer	Molina	Richards	Washburn
Calloway	Gandhi	Morgan	Rider	Zabidi
Delluci	Garcia	Murphy	Rodriguez	Zhao

12. **Apartment living** You are planning a report on apartment living in a college town. You decide to select three apartment complexes at random for in-depth interviews with residents.

(a) Explain how you would use a line of Table D to choose an SRS of 3 complexes from the list below. Explain your method clearly enough for a classmate to obtain your results.

(b) Use line 117 to select the sample. Show how you use each of the digits.

Ashley Oaks	Chauncey Village	Franklin Park	Richfield
Bay Pointe	Country Squire	Georgetown	Sagamore Ridge
Beau Jardin	Country View	Greenacres	Salem Courthouse
Bluffs	Country Villa	Lahr House	Village Manor
Brandon Place	Crestview	Mayfair Village	Waterford Court
Briarwood	Del-Lynn	Nobb Hill	Williamsburg
Brownstone	Fairington	Pemberly Courts	
Burberry	Fairway Knolls	Peppermill	
Cambridge	Fowler	Pheasant Run	

13. **Sampling the forest** To gather data on a 1200-acre pine forest in Louisiana, the U.S. Forest Service laid a grid of 1410 equally spaced circular plots over a map of the forest. A ground survey visited a sample of 10% of these plots.<sup>12</sup>

(a) Explain how you would use technology or Table D to choose an SRS of 141 plots. Your description should be clear enough for a classmate to obtain your results.

(b) Use your method from (a) to choose the first 3 plots.

14. **Sampling gravestones** The local genealogical society in Coles County, Illinois, has compiled records on all 55,914 gravestones in cemeteries in the county for the years 1825 to 1985. Historians plan to use these records to learn about African Americans in Coles County's history. They first choose an SRS of 395 records to check their accuracy by visiting the actual gravestones.<sup>13</sup>

(a) Explain how you would use technology or Table D to choose the SRS. Your description should be clear enough for a classmate to obtain your results.

(b) Use your method from (a) to choose the first 3 gravestones.

15. **Random digits** In using Table D repeatedly to choose random samples, you should not always begin at the same place, such as line 101. Why not?

16. **Random digits** Which of the following statements are true of a table of random digits, and which are false? Briefly explain your answers.

- (a) There are exactly four 0s in each row of 40 digits.  
 (b) Each pair of digits has chance 1/100 of being 00.  
 (c) The digits 0000 can never appear as a group, because this pattern is not random.

17. **iPhones** Suppose 1000 iPhones are produced at a factory today. Management would like to ensure that the phones' display screens meet their quality standards before shipping them to retail stores. Since it takes about 10 minutes to inspect an individual phone's display screen, managers decide to inspect a sample of 20 phones from the day's production.

(a) Explain why it would be difficult for managers to inspect an SRS of 20 iPhones that are produced today.

(b) An eager employee suggests that it would be easy to inspect the last 20 iPhones that were produced today. Why isn't this a good idea?

(c) Another employee recommends inspecting every fiftieth iPhone that is produced. Explain carefully why this sampling method is *not* an SRS.

18. **Dead trees** On the west side of Rocky Mountain National Park, many mature pine trees are dying due to infestation by pine beetles. Scientists would like to use sampling to estimate the proportion of all pine trees in the area that have been infected.

(a) Explain why it wouldn't be practical for scientists to obtain an SRS in this setting.

(b) A possible alternative would be to use every pine tree along the park's main road as a sample. Why is this sampling method biased?

(c) Suppose that a more complicated random sampling plan is carried out, and that 35% of the pine trees in the sample are infested by the pine beetle. Can scientists conclude that 35% of *all* the pine trees on the west side of the park are infested? Why or why not?

19. **Who goes to the convention?** A club has 30 student members and 10 faculty members. The students are

Abel	Fisher	Huber	Miranda	Reinmann
Carson	Ghosh	Jimenez	Moskowitz	Santos
Chen	Griswold	Jones	Neyman	Shaw
David	Hein	Kim	O'Brien	Thompson
Deming	Hernandez	Klotz	Pearl	Utts
Elashoff	Holland	Liu	Potter	Varga

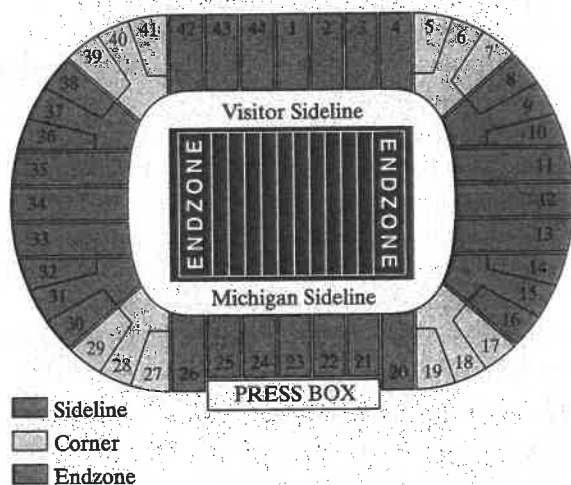
The faculty members are

Andrews	Fernandez	Kim	Moore	West
Besicovitch	Gupta	Lightman	Phillips	Yang

The club can send 4 students and 2 faculty members to a convention. It decides to choose those who will go by random selection. How will you label the two strata? Use Table D, beginning at line 123, to choose a stratified random sample of 4 students and 2 faculty members.

20. **Sampling by accountants** Accountants often use stratified samples during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are in amounts over \$50,000; 500 are in amounts between \$1000 and \$50,000; and the remaining 4400 are in amounts under \$1000. Using these groups as strata, you decide to verify all the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. How would you label the two strata from which you will sample? Use Table D, starting at line 115, to select only the first 3 accounts from each of these strata.

21. **Go Blue!** Michigan Stadium, also known as "The Big House," seats over 100,000 fans for a football game. The University of Michigan athletic department plans to conduct a survey about concessions that are sold during games. Tickets are most expensive for seats near the field and on the sideline. The cheapest seats are high up in the end zones (where one of the authors sat as a student). A map of the stadium is shown.



- (a) The athletic department is considering a stratified random sample. What would you recommend as the strata? Why?
- (b) Explain why a cluster sample might be easier to obtain. What would you recommend for the clusters? Why?
22. **How was your stay?** A hotel has 30 floors with 40 rooms per floor. The rooms on one side of the hotel

face the water, while rooms on the other side face a golf course. There is an extra charge for the rooms with a water view. The hotel manager wants to survey 120 guests who stayed at the hotel during a convention about their overall satisfaction with the property.

- (a) Explain why choosing a stratified random sample might be preferable to an SRS in this case. What would you use as strata?
- (b) Why might a cluster sample be a simpler option? What would you use as clusters?
23. **Is it an SRS?** A corporation employs 2000 male and 500 female engineers. A stratified random sample of 200 male and 50 female engineers gives each engineer 1 chance in 10 to be chosen. This sample design gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.
24. **Attitudes toward alcohol** At a party there are 30 students over age 21 and 20 students under age 21. You choose at random 3 of those over 21 and separately choose at random 2 of those under 21 to interview about attitudes toward alcohol. You have given every student at the party the same chance to be interviewed: what is the chance? Why is your sample not an SRS?
25. **High-speed Internet** Laying fiber-optic cable is expensive. Cable companies want to make sure that, if they extend their lines out to less dense suburban or rural areas, there will be sufficient demand and the work will be cost-effective. They decide to conduct a survey to determine the proportion of households in a rural subdivision that would buy the service. They select a simple random sample of 5 blocks in the subdivision and survey each family that lives on one of those blocks.
- (a) What is the name for this kind of sampling method?
- (b) Suppose there are 65 blocks in the subdivision. Use technology or Table D to select 5 blocks to be sampled. Explain your method clearly.

26. **Systematic random sample** Sample surveys often use a systematic random sample to choose a sample of apartments in a large building or housing units in a block at the last stage of a multistage sample. Here is a description of how to choose a systematic random sample.



Suppose that we must choose 4 addresses out of 100. Because  $100/4 = 25$ , we can think of the list as four lists of 25 addresses. Choose 1 of the first 25 addresses at random using Table D. The sample contains this address and the addresses 25, 50, and 75 places down the list from it. If the table gives 13, for example, then the systematic random sample consists of the addresses numbered 13, 38, 63, and 88.

- (a) Use Table D to choose a systematic random sample of 5 addresses from a list of 200. Enter the table at line 120.
- (b) Like an SRS, a systematic random sample gives all individuals the same chance to be chosen. Explain why this is true. Then explain carefully why a systematic sample is not an SRS.
27. **Sampling frame** Ideally, the sampling frame in a sample survey should list every individual in the population, but in practice, this is often difficult. Suppose that a sample of households in a community is selected at random from the telephone directory. Explain how this sampling method results in undercoverage that could lead to bias.
28. **Sampling frame** Refer to the previous exercise. It is more common in telephone surveys to use random digit dialing equipment that selects the last four digits of a telephone number at random after being given the exchange (the first three digits). Explain how this sampling method results in undercoverage that could lead to bias.
29. **Baseball tickets** Suppose you want to know the average amount of money spent by the fans attending opening day for the Cleveland Indians baseball season. You get permission from the team's management to conduct a survey at the stadium, but they will not allow you to bother the fans in the club seating or box seats (the most expensive seating). Using a computer, you randomly select 500 seats from the rest of the stadium. During the game, you ask the fans in those seats how much they spent that day.
- (a) Provide a reason why this survey might yield a biased result.
- (b) Explain whether the reason you provided in (a) is a sampling error or a nonsampling error.
30. **What kind of error?** Which of the following are sources of *sampling error* and which are sources of *nonsampling error*? Explain your answers.
- (a) The subject lies about past drug use.
- (b) A typing error is made in recording the data.
- (c) Data are gathered by asking people to mail in a coupon printed in a newspaper.
31. **Nonresponse** A survey of drivers began by randomly sampling all listed residential telephone numbers in the United States. Of 45,956 calls to these numbers, 5029 were completed. The goal of the survey was to estimate how far people drive, on average, per day.<sup>14</sup>
- (a) What was the rate of nonresponse for this sample?
- (b) Explain how nonresponse can lead to bias in this survey. Be sure to give the direction of the bias.
32. **Ring-no-answer** A common form of nonresponse in telephone surveys is "ring-no-answer." That is, a call is made to an active number but no one answers. The Italian National Statistical Institute looked at nonresponse to a government survey of households in Italy during the periods January 1 to Easter and July 1 to August 31. All calls were made between 7 and 10 P.M., but 21.4% gave "ring-no-answer" in one period versus 41.5% "ring-no-answer" in the other period.<sup>15</sup> Which period do you think had the higher rate of no answers? Why? Explain why a high rate of nonresponse makes sample results less reliable.
33. **Running red lights** The sample described in Exercise 31 produced a list of 5024 licensed drivers. The investigators then chose an SRS of 880 of these drivers to answer questions about their driving habits. One question asked was: "Recalling the last ten traffic lights you drove through, how many of them were red when you entered the intersections?" Of the 880 respondents, 171 admitted that at least one light had been red. A practical problem with this survey is that people may not give truthful answers. What is the likely direction of the bias: do you think more or fewer than 171 of the 880 respondents really ran a red light? Why?
34. **Seat belt use** A study in El Paso, Texas, looked at seat belt use by drivers. Drivers were observed at randomly chosen convenience stores. After they left their cars, they were invited to answer questions that included questions about seat belt use. In all, 75% said they always used seat belts, yet only 61.5% were wearing seat belts when they pulled into the store parking lots.<sup>16</sup> Explain the reason for the bias observed in responses to the survey. Do you expect bias in the same direction in most surveys about seat belt use?
35. **Wording bias** Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?
- (a) "Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?"
- (b) "Do you agree that a national system of health insurance should be favored because it would provide health insurance for everyone and would reduce administrative costs?"
- (c) "In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?"
36. **Checking for bias** Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?
- (a) Which of the following best represents your opinion on gun control?
1. The government should confiscate our guns.
  2. We have the right to keep and bear arms.



(b) A freeze in nuclear weapons should be favored because it would begin a much-needed process to stop everyone in the world from building nuclear weapons now and reduce the possibility of nuclear war in the future. Do you agree or disagree?

**Multiple choice:** Select the best answer for Exercises 37 to 42.

37. The Web portal AOL places opinion poll questions next to many of its news stories. Simply click your response to join the sample. One of the questions in January 2008 was "Do you plan to diet this year?" More than 30,000 people responded, with 68% saying "Yes." You can conclude that
- about 68% of Americans planned to diet in 2008.
  - the poll used a convenience sample, so the results tell us little about the population of all adults.
  - the poll uses voluntary response, so the results tell us little about the population of all adults.
  - the sample is too small to draw any conclusion.
  - None of these.
38. Archaeologists plan to examine a sample of 2-meter-square plots near an ancient Greek city for artifacts visible in the ground. They choose separate random samples of plots from floodplain, coast, foothills, and high hills. What kind of sample is this?
- A cluster sample
  - A convenience sample
  - A simple random sample
  - A stratified random sample
  - A voluntary response sample
39. Your statistics class has 30 students. You want to call an SRS of 5 students from your class to ask where they use a computer for the online exercises. You label the students 01, 02, ..., 30. You enter the table of random digits at this line:
- 14459 26056 31424 80371 65103 62253 22490 61181
- Your SRS contains the students labeled
- 14, 45, 92, 60, 56.
  - 14, 31, 03, 10, 22.
  - 14, 03, 10, 22, 22.
  - 14, 03, 10, 22, 06.
  - 14, 03, 10, 22, 11.
40. When we take a census, we attempt to collect data from
- a stratified random sample.
  - every individual selected in an SRS.
  - every individual in the population.
  - a voluntary response sample.
  - a convenience sample.
41. An example of a nonsampling error that can reduce the accuracy of a sample survey is
- using voluntary response to choose the sample.
  - using the telephone directory as the sampling frame.
  - interviewing people at shopping malls to obtain a sample.
  - variation due to chance in choosing a sample at random.
  - inability to contact many members of the sample.
42. A simple random sample of 1200 adult Americans is selected, and each person is asked the following question: "In light of the huge national deficit, should the government at this time spend additional money to establish a national system of health insurance?" Only 39% of those responding answered "Yes." This survey
- is reasonably accurate since it used a large simple random sample.
  - needs to be larger since only about 24 people were drawn from each state.
  - probably understates the percent of people who favor a system of national health insurance.
  - is very inaccurate but neither understates nor overstates the percent of people who favor a system of national health insurance. Since simple random sampling was used, it is unbiased.
  - probably overstates the percent of people who favor a system of national health insurance.
43.  **Sleep debt (3.2)** A researcher reported that the average teenager needs 9.3 hours of sleep per night but gets only 6.3 hours.<sup>17</sup> By the end of a 5-day school week, a teenager would accumulate about 15 hours of "sleep debt." Students in a high school statistics class were skeptical, so they gathered data on the amount of sleep debt (in hours) accumulated over time (in days) by a random sample of 25 high school students. The resulting least-squares regression equation for their data is  $\text{Sleep debt} = 2.23 + 3.17(\text{days})$ . Do the students have reason to be skeptical of the research study's reported results? Explain.
44.  **Internet charges (2.1)** Some Internet service providers (ISPs) charge companies based on how much bandwidth they use in a month. One method that ISPs use for calculating bandwidth is to find the 95th percentile of a company's usage based on samples of hundreds of 5-minute intervals during a month.
- Explain what "95th percentile" means in this setting.
  - Which would cost a company more: the 95th percentile method or a similar approach using the 98th percentile? Justify your answer.

## 4.2

## Experiments

In Section 4.2,  
you'll learn about:

- Observational study versus experiment
- The language of experiments
- How to experiment badly
- How to experiment well: The randomized comparative experiment
- Three principles of experimental design
- Experiments: What can go wrong?
- Inference for experiments
- Blocking
- Matched pairs design

A sample survey aims to gather information about a population without disturbing the population in the process. Sample surveys are one kind of **observational study**. Other observational studies watch the behavior of animals in the wild or the interactions between teacher and students in the classroom. This section is about statistical designs for **experiments**, a very different way to produce data.

## Observational Study versus Experiment

In contrast to observational studies, experiments don't just observe individuals or ask them questions. They actively impose some *treatment* in order to measure the response. Experiments can answer questions like "Does aspirin reduce the chance of a heart attack?" and "Can yoga help dogs live longer?"

### DEFINITION: Observational study and experiment

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.

An **experiment** deliberately imposes some treatment on individuals to measure their responses.

The goal of an observational study can be to describe some group or situation, to compare groups, or to examine relationships between variables. The purpose of an experiment is to determine whether the treatment causes a change in the response. An observational study, even one based on a random sample, is a poor way to gauge the effect that changes in one variable have on another variable. To see the response to a change, we must actually impose the change. *When our goal is to understand cause and effect, experiments are the only source of fully convincing data.* For this reason, the distinction between observational study and experiment is one of the most important in statistics.

### EXAMPLE

#### *Does Taking Hormones Reduce Heart Attack Risk after Menopause?*

##### Observation versus experiment

Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said "Yes." Women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits.

The evidence in favor of hormone replacement came from a number of *observational studies* that compared women who were taking hormones with others who were not. But the women who chose to take hormones were richer and better educated and saw doctors more often than women who didn't take hormones. Because the



women who took hormones did many other things to maintain their health, it isn't surprising that they had fewer heart attacks.

To get convincing data on the link between hormone replacement and heart attacks, we should do an *experiment*. Experiments don't let women decide what to do. They assign women to either hormone replacement or to dummy pills that look and taste the same as the hormone pills. The assignment is done by a coin toss, so that all kinds of women are equally likely to get either treatment. By 2002, several experiments with women of different ages agreed that hormone replacement does *not* reduce the risk of heart attacks. The National Institutes of Health, after reviewing the evidence, concluded that the first studies were wrong. Taking hormones after menopause quickly fell out of favor.<sup>18</sup>

From Chapter 3: A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or influence changes in a response variable.

For each of these studies, the *explanatory variable* was whether a woman took hormones, and the *response variable* was whether the woman had a heart attack. Researchers wanted to argue that changes in the explanatory variable (taking hormones or not) actually caused changes in the response variable (having a heart attack or not). In the early observational studies, however, the effect of taking hormones was mixed up with the characteristics of women who chose to take them. These characteristics are **lurking variables** that make it hard to see the true relationship between the explanatory and response variables.

#### DEFINITION: Lurking variable

A **lurking variable** is a variable that is not among the explanatory or response variables in a study but that may influence the response variable.

Let's consider two lurking variables from the observational studies of hormone replacement: number of doctor visits per year and age. The women who chose to take hormones visited their doctors more often than the women who didn't take hormones. Did the women in the hormone group have fewer heart attacks because they got better health care or because they took hormones? We can't be sure. A situation like this, in which the effects of two variables on a response variable cannot be separated from each other, is called **confounding**.

What about age? Older women are at greater risk of having a heart attack than younger women. If the women who took hormones were generally younger than those who didn't, we'd have more confounding. That wasn't the case, however. There was no link between age and group membership (hormones or not) in the observational studies. *With no association between the lurking variable and the explanatory variable, there can be no confounding.*

Some people call a lurking variable that results in confounding, like the number of doctor visits per year in this case, a *confounding variable*.

#### DEFINITION: Confounding

**Confounding** occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other.

**AP EXAM TIP** If you are asked to identify a possible confounding variable in a given setting, you are expected to explain how the variable you choose (1) is associated with the explanatory variable and (2) affects the response variable.

*Observational studies of the effect of one variable on another often fail because of confounding between the explanatory variable and one or more lurking variables.* Well-designed experiments take steps to prevent confounding. The later hormone therapy experiments avoided confounding by letting chance decide who took hormones and who didn't. That way, women who took better care of themselves were split about evenly between the two groups. So were older women and younger women. When these experiments found no reduction in heart attack risk for women taking hormones, researchers began to doubt the results of the earlier observational studies. The moral of the story is simple: *beware the lurking variable!*



## CHECK YOUR UNDERSTANDING

1. Does reducing screen brightness increase battery life in laptop computers? To find out, researchers obtained 30 new laptops of the same brand. They chose 15 of the computers at random and adjusted their screens to the brightest setting. The other 15 laptop screens were left at the default setting—moderate brightness. Researchers then measured how long each machine's battery lasted. Was this an observational study or an experiment? Justify your answer.

Questions 2 to 4 refer to the following setting. Does eating dinner with their families improve students' academic performance? According to an ABC News article, "Teenagers who eat with their families at least five times a week are more likely to get better grades in school."<sup>19</sup> This finding was based on a sample survey conducted by researchers at Columbia University.

2. Was this an observational study or an experiment? Justify your answer.

3. What are the explanatory and response variables?

4. Explain clearly why such a study cannot establish a cause-and-effect relationship. Suggest a lurking variable that may be confounded with whether families eat dinner together.



## The Language of Experiments

An experiment is a statistical study in which we actually do something (a **treatment**) to people, animals, or objects (the **experimental units**) to observe the response. Here is the basic vocabulary of experiments.

### DEFINITION: Treatment, experimental units, subjects

A specific condition applied to the individuals in an experiment is called a **treatment**. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

The **experimental units** are the smallest collection of individuals to which treatments are applied. When the units are human beings, they often are called **subjects**.

The best way to learn the language of experiments is to practice using it.



**EXAMPLE***When Will I Ever Use This Stuff?***Vocabulary of experiments**

Researchers at the University of North Carolina were concerned about the increasing dropout rate in the state's high schools, especially for low-income students. Surveys of recent dropouts revealed that many of these students had started to lose interest during middle school. They said they saw little connection between what they were studying in school and their future plans. To change this perception, researchers developed a program called CareerStart. The big idea of the program is that teachers show students how the topics they learn get used in specific careers.

To test the effectiveness of CareerStart, the researchers recruited 14 middle schools in Forsyth County to participate in an experiment. Seven of the schools, chosen at random, used CareerStart along with the district's standard curriculum. The other seven schools just followed the standard curriculum. Researchers followed both groups of students for several years, collecting data on students' attendance, behavior, standardized test scores, level of engagement in school, and whether the students graduated from high school. *Results:* Students at schools that used CareerStart generally had better attendance and fewer discipline problems, earned higher test scores, reported greater engagement in their classes, and were more likely to graduate.<sup>20</sup>

**PROBLEM:** Identify the experimental units, explanatory and response variables, and the treatments in the CareerStart experiment.

**SOLUTION:** The *experimental units* are 14 middle schools in Forsyth County, NC. The *explanatory variable* is whether the school used the CareerStart program with its students. This experiment compares two treatments: (1) the standard middle school curriculum and (2) the standard curriculum plus CareerStart. Several *response variables* were measured, including test scores, attendance, behavior, student engagement, and graduation rates.

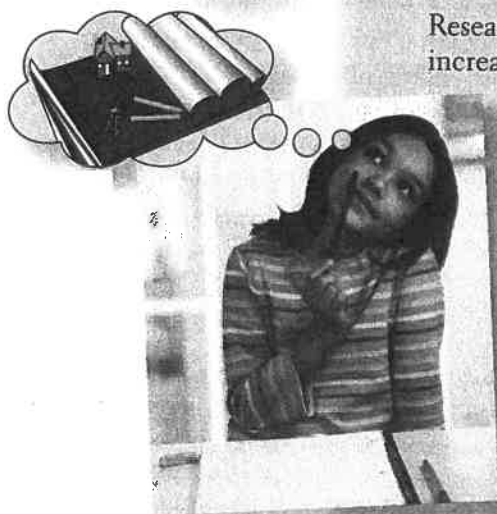
**For Practice Try Exercise 51**

The previous example illustrates the big advantage of experiments over observational studies: *experiments can give good evidence for causation*. In an experiment, we study the effects of the specific treatments we are interested in, while trying to control for the effect of lurking variables. For instance, the students in all 14 schools followed the standard curriculum. To ensure that the two groups of schools were as similar as possible before the treatments were administered, researchers let chance decide which 7 schools would use CareerStart. The only systematic difference between the schools was the educational treatment. When students from the CareerStart schools did better, we can be confident that the program made the difference.

Sometimes, the explanatory variables in an experiment are called **factors**. Many experiments study the joint effects of several factors. In such an experiment, each treatment is formed by combining a specific value (often called a **level**) of each of the factors. Here's an example of a multifactor experiment.

Factors

Level



Note that the experimental units in the CareerStart example are the schools, not individual students. Experimental units are the smallest collection of individuals to which treatments are applied. The curricular treatments were administered to entire schools, so those are the experimental units.



**EXAMPLE****TV Advertising****Experiments with multiple explanatory variables**

What are the effects of repeated exposure to an advertising message? The answer may depend on both the length of the ad and on how often it is repeated. An experiment investigated this question using undergraduate students as subjects. All subjects viewed a 40-minute television program that included ads for a digital camera. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either 1, 3, or 5 times during the program. After viewing, all the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it.<sup>21</sup>

**PROBLEM:** For the advertising study,

- Identify the explanatory and response variables, and
- list all the treatments.

**SOLUTION:**

(a) This experiment has 2 explanatory variables (factors): length of the commercial and number of repetitions. The response variables include measures of subjects' recall of the ad, their attitudes about the digital camera, and whether they intend to purchase it.

(b) There are 2 different lengths of commercial (30 and 90 seconds) and three different numbers of repetitions (1, 3, and 5). The 6 combinations consisting of one level of each factor form the 6 treatments shown in Figure 4.3: (1) 30 seconds, 1 time; (2) 30 seconds, 3 times; (3) 30 seconds, 5 times; (4) 90 seconds, 1 time; (5) 90 seconds, 3 times; (6) 90 seconds, 5 times.

		Factor B: Repetitions		
		1 time	3 times	5 times
Factor A: Length	30 seconds	1	2	3
	90 seconds	4	5	6

Subjects assigned to Treatment 3 see a 30-second ad five times during the program.

**FIGURE 4.3** The six treatments in the TV ad experiment. Combinations of values of the two explanatory variables (factors) form six treatments.

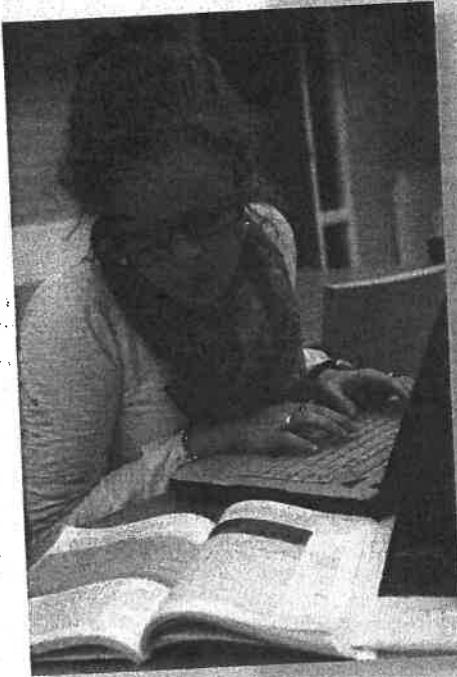
**For Practice Try Exercise 55**

This example shows how experiments allow us to study the combined effect of several factors. The interaction of several factors can produce effects that could not be predicted from looking at the effect of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest. But if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. The two-factor experiment in the TV advertising example will help us find out.

## How to Experiment Badly

Experiments are the preferred method for examining the effect of one variable on another. By imposing the specific treatment of interest and controlling other influences, we can pin down cause and effect. Good designs are essential for effective experiments, just as they are for sampling. To see why, let's start with an example of a bad experimental design.

## EXAMPLE

*Which Works Better: Online or In-Class SAT Preparation?***A bad experiment**

A high school regularly offers a review course to prepare students for the SAT. This year, budget cuts will allow the school to offer only an online version of the course. Over the past 10 years, the average SAT score of students in the classroom course was 1620. The online group gets an average score of 1780. That's roughly 10% higher than the long-time average for those who took the classroom review course. Is the online course more effective?

This experiment has a very simple design. A group of subjects (the students) were exposed to a treatment (the online course), and the outcome (SAT scores) was observed. Here is the design:

Students → Online course → SAT scores

A closer look showed that the students in the online review course were quite different from the students who took the classroom course in past years. They had higher GPAs and were taking more AP classes.

The effect of online versus in-class instruction is mixed up with the effect of these lurking variables. Maybe the online students earned higher SAT scores because they were smarter to begin with, not because the online course prepared them better. This confounding

prevents us from concluding that the online course is more effective than classroom instruction.



Many laboratory experiments use a design like the one in the online SAT course example:

Experimental units → Treatment → Measure response

In the lab environment, simple designs often work well. Field experiments and experiments with animals or people deal with more variable conditions. *Outside the lab, badly designed experiments often yield worthless results because of confounding.*

## How to Experiment Well: The Randomized Comparative Experiment

The remedy for the confounding in the SAT prep course example is to do a *comparative experiment* in which some students are taught in the classroom and other, similar students take the course online. Most well-designed experiments compare two or more treatments.

Comparison alone isn't enough to produce results we can trust. If the treatments are given to groups that differ greatly when the experiment begins, *bias* will result. For example, if we allow students to select online or classroom instruction,

more self-motivated students are likely to sign up for the online course. Allowing personal choice will bias our results in the same way that volunteers bias the results of online opinion polls. The solution to the problem of bias in sampling is random selection. In experiments, the solution is **random assignment**.

### DEFINITION: Random assignment

In an experiment, **random assignment** means that experimental units are assigned to treatments at random, that is, using some sort of chance process.

Let's look at how random assignment can be used to improve the design of the SAT prep course experiment.

## EXAMPLE

### *SAT Prep: Online versus Classroom*

#### How random assignment works



This year, the high school has enough budget money to compare the online SAT course with traditional classroom instruction. Fifty students have agreed to participate in an experiment comparing the two instructional methods.

**PROBLEM:** Describe how you would randomly assign 25 students to each of the two methods:

- (a) Using 50 identical slips of paper
- (b) Using Table D
- (c) Using technology

**SOLUTION:**

(a) The simplest way would be to use the "hat method." Write each subject's name on one of the slips. Put all the slips in a hat and mix them thoroughly. Draw them out one at a time until you have 25 slips. These 25 students will take the online course. The remaining 25 students will take the classroom course. Alternatively, you could write "online" on 25 of the slips and "classroom" on the other 25 slips. Then put the slips in

a hat and mix them well. Have students come up one by one and (without looking) pick a slip from the hat. This guarantees 25 students per group, with the treatments assigned by chance.

(b) We can use the two-step process from random sampling to do the random assignment.

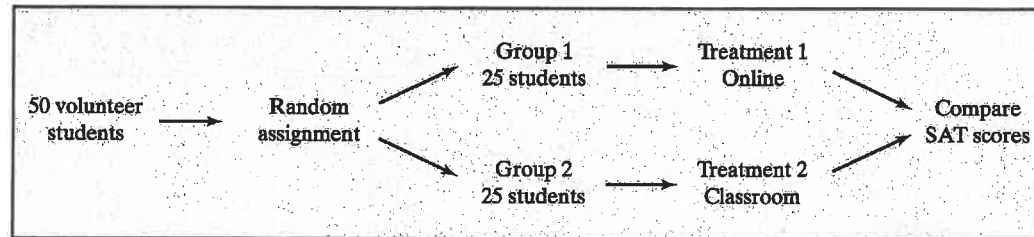
**Step 1: Label.** Give labels 01, 02, 03, ..., 49, 50 to the subjects in alphabetical order by last name.

**Step 2: Table.** Go to a line of Table D and read two-digit groups moving from left to right. The first 25 distinct labels between 01 and 50 that you encounter select the online group. The remaining 25 students will take the classroom course. Ignore repeated labels and groups of digits from 51 to 00.

(c) Give numbers 1, 2, 3, ..., 49, 50 to the subjects in alphabetical order by last name. Then use your calculator's `randInt` command or a computer's random number generator to produce numbers between 1 and 50. The first 25 different numbers chosen select the students for the online course. The remaining 25 subjects will take the classroom course.

For Practice Try Exercise 59

The experimental design in this example is *comparative* because it compares two treatments (the two instructional settings). It is *randomized* because the subjects are assigned to the treatments by chance. The diagram in Figure 4.4 presents the details: random assignment, the sizes of the groups and which treatment they receive, and the response variable. There are, as we will see later, statistical reasons for using treatment groups that are about equal in size. This type of design is called a **completely randomized design**.



**FIGURE 4.4** Outline of a completely randomized design to compare online and classroom instruction.

#### DEFINITION: Completely randomized design

In a **completely randomized design**, the treatments are assigned to all the experimental units completely by chance.

Notice that the definition of a completely randomized design does not require that each treatment be assigned to an equal number of experimental units. It does specify that the assignment of treatments must occur completely at random.

### THINK ABOUT IT

Does using chance to assign treatments in an experiment guarantee a completely randomized design? Actually, no. Let's return to the SAT prep course experiment involving 50 students. Another way to randomly assign the subjects to the two treatments is by tossing a coin. Have each student come forward and toss a coin. If it's heads, then the student will take the course online. If it's tails, then the student will take the classroom course.

As long as all 50 students toss a coin, this is still a completely randomized design. Of course, the two experimental groups are unlikely to contain exactly 25 students each due to the chance variation in coin tosses.

The problem comes if we try to force the two groups to have equal sizes. Suppose we let the coin tossing continue until one of the groups has 25 students and then place the remaining students in the other group. This is no longer a completely randomized design, because the last few students aren't being assigned to one of the treatments by chance. In fact, these students will all end up in the same group, which could lead to bias if these individuals share some characteristic that would systematically affect the response variable. For example, if the students came to toss the coin last because they're lazier than the other students who volunteered, then the SAT prep class that they're in will seem less effective than it really is.

Completely randomized designs can compare any number of treatments. Here is an experiment that compares three treatments.



## EXAMPLE

## Conserving Energy

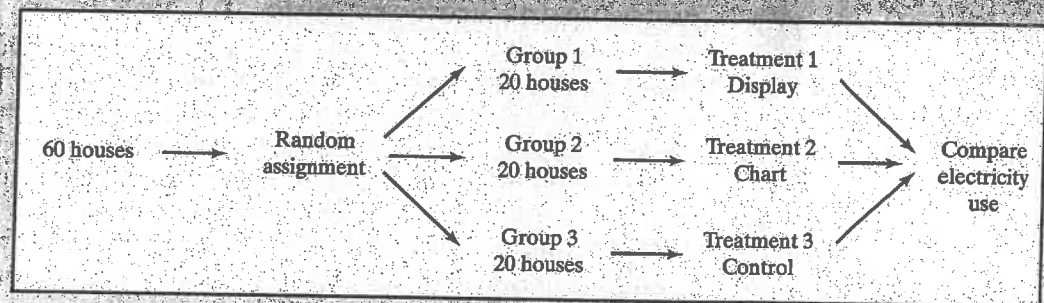
## A completely randomized design

Many utility companies have introduced programs to encourage energy conservation among their customers. An electric company considers placing small digital displays in households to show current electricity use and what the cost would be if this use continued for a month. Will the displays reduce electricity use? One cheaper approach is to give customers a chart and information about monitoring their electricity use from their outside meter. Would this method work almost as well? The company decides to conduct an experiment to compare these two approaches (display, chart) with a *control group* of customers who receive information about energy consumption but no help in monitoring electricity use.

**PROBLEM:** Outline a completely randomized design involving 60 single-family residences in the same city who are willing to participate in such an experiment. Write a few sentences describing how you would implement your design.

**SOLUTION:** Figure 4.5 outlines the design. We'll randomly assign 20 houses to each of three treatments: digital display, chart plus information, and control. Our response variable is the total amount of electricity used in a year.

**AP EXAM TIP** If you are asked to describe the design of an experiment on the AP exam, you won't get full credit for a diagram like Figure 4.5. You are expected to describe how the treatments are assigned to the experimental units and to clearly state what will be measured or compared. Some students prefer to start with a diagram and then add a few sentences. Others choose to skip the diagram and put their entire response in narrative form.



**FIGURE 4.5** Outline of a completely randomized design to compare three energy-saving programs.

To implement the design, start by labeling each household with a distinct number from 1 to 60. Write the labels on 60 identical slips of paper, put them in a hat, and mix them well. Draw out 20 slips. The corresponding households will be given digital displays showing current electricity use. Now draw out 20 more slips. Those houses will use a chart. The remaining 20 houses will be given information about energy consumption but no way to monitor their usage. At the end of the year, compare how much electricity was used by the homes in the three groups.

For Practice Try Exercise 61

Control group

Why did we include a **control group** of 20 houses in the energy conservation experiment? The primary purpose of a control group is to provide a baseline for comparing the effects of the other treatments. Without such a comparison group, we wouldn't be able to tell whether homes with digital displays or charts used less electricity than homes without such aids.





## THINK ABOUT IT

Was a control group really necessary? You might be thinking that the change in electricity use from last year to this year in the houses with displays and charts would tell us whether these treatments helped. Unfortunately, it's not that simple. Suppose last year's temperatures were more extreme than this year's. Then many households might show a decrease in electricity use, but we couldn't be sure whether this change was due to the weather or to the treatments. (Can you say confounding?!) .....

Many experiments (like the one in the previous example) include a control group that receives an inactive treatment. However, a control group can be given an active treatment. Suppose we want to compare the effectiveness of a newly developed drug for treating a serious disease with a drug that's already known to work. In that case, the experimental units that receive the existing drug form the control group.

Some experimental designs don't include a control group. That's appropriate if researchers simply want to compare the effects of several treatments, and not to determine whether any of them works better than an inactive treatment. For instance, maybe a state's highway department wants to see which of three brands of paint will last longest when marking lane lines on the freeway.



### CHECK YOUR UNDERSTANDING

Music students often don't evaluate their own performances accurately. Can small-group discussions help? The subjects were 29 students preparing for the end-of-semester performance that is an important part of their grade. Assign 15 students to the treatment: videotape a practice performance, ask the student to evaluate it, then have the student discuss the tape with a small group of other students. The remaining 14 students form a control group who watch and evaluate their tapes alone. At the end of the semester, the discussion-group students evaluated their final performance more accurately.<sup>22</sup>

1. Outline a completely randomized design for this experiment. Follow the model of Figure 4.4.
2. Describe how you would carry out the random assignment. Provide enough detail that a classmate could implement your procedure.
3. What is the purpose of the control group in this experiment?

## Three Principles of Experimental Design

Randomized comparative experiments are designed to give good evidence that differences in the treatments actually *cause* the differences we see in the response. The logic is as follows:

- A proper comparative design ensures that influences other than the experimental treatments operate equally on all groups. This helps control for the effects of lurking variables.
- Random assignment helps balance out the effects of lurking variables that we can't control, or don't think of, on the treatment groups. That is, random assignment forms groups of experimental units that should be similar in all respects before the treatments are applied.

- Therefore, since the groups are roughly equivalent except for the treatments, any differences in average response must be due either to the treatments or to the play of chance in the random assignment of experimental units to the treatments.

The first two bullets give us two basic principles of experimental design: **control** and **random assignment**. That “either-or” in the last bullet deserves some more thought. In the SAT prep course example, we can’t say that *any* difference between the average SAT scores of students enrolled online and in the classroom must be caused by a difference in the effectiveness of the two types of instruction. There would be *some* difference even if both groups received the same instruction, because of variation among students in background and study habits. Chance assigns students to one group or the other, and this creates a chance difference between the groups.

We would not trust an experiment with just one student in each group, for example. The results would depend too much on which group got lucky and received the stronger student. If we assign many subjects to each group, however, the effects of chance will average out, and there will be little difference in the average responses in the two groups unless the treatments themselves cause a difference. Our third design principle is **replication**: use enough experimental units to distinguish a difference in the effects of the treatments from chance variation.

### Replication

In statistics, replication means “use enough subjects.” In other fields, the term “replication” has a different meaning. If one experiment is conducted and then the same or a similar experiment is independently conducted in a different location by different investigators, this is known as replication. That is, in this case, replication means repeatability.

### Summary: Principles of Experimental Design

The basic principles for designing experiments are as follows:

1. **Control** for lurking variables that might affect the response: Use a comparative design and ensure that the only systematic difference between the groups is the treatment administered.
2. **Random assignment**: Use impersonal chance to assign experimental units to treatments. This helps create roughly equivalent groups of experimental units by balancing the effects of lurking variables that aren’t controlled on the treatment groups.
3. **Replication**: Use enough experimental units in each group so that any differences in the effects of the treatments can be distinguished from chance differences between the groups.

Let’s see how these principles were used in designing a famous medical experiment.

### EXAMPLE

#### *The Physicians’ Health Study*

#### A well-designed experiment

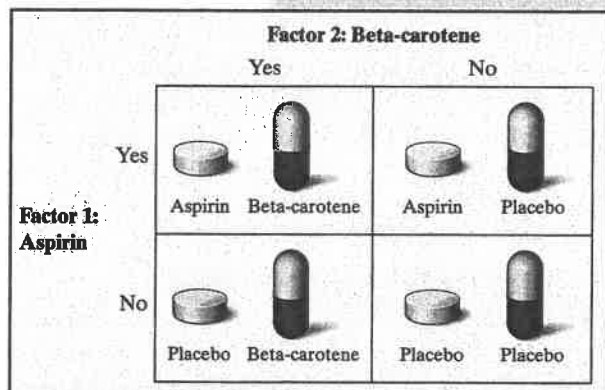
Does regularly taking aspirin help protect people against heart attacks? The Physicians’ Health Study was a medical experiment that helped answer this question. In fact, the Physicians’ Health Study looked at the effects of two drugs: aspirin and beta-carotene. Researchers wondered whether beta-carotene would help prevent some forms of cancer. The subjects in this experiment were 21,996 male physicians. There were two explanatory variables (factors), each having two levels: aspirin (yes or no) and beta-carotene (yes or no). Combinations of



Placebo

the levels of these factors form the four treatments shown in Figure 4.6. One-fourth of the subjects were assigned at random to each of these treatments.

On odd-numbered days, the subjects took either a tablet that contained aspirin or a dummy pill that looked and tasted like the aspirin but had no active ingredient (a **placebo**). On even-numbered days, they took either a capsule containing beta-carotene or a placebo. There were several response variables—the study looked for heart attacks, several kinds of cancer, and other medical outcomes. After several years, 239 of the placebo group but only 139 of the aspirin group had suffered heart attacks. This difference is large enough to give good evidence that taking aspirin does reduce heart attacks.<sup>23</sup> It did not appear, however, that beta-carotene had any effect on preventing cancer.



**FIGURE 4.6** The treatments in the Physicians' Health Study.

**PROBLEM:** Explain how each of the three principles of experimental design was used in the Physicians' Health Study.

**SOLUTION:** Researchers attempted to *control* for the effects of lurking variables by using a design that compared both of the active treatments to a placebo and by having all subjects follow the same schedule of pill taking. *Random assignment* was used to determine which subjects received each of the four treatment combinations. This helped ensure that the treatment groups were roughly equivalent to begin with. There were over 5000 subjects per treatment group. This *replication* helped ensure that any sizable difference in response among the groups was due to the treatments and not to chance variation in the random assignment.

For Practice Try Exercise 67

Why did researchers decide to do the Physicians' Health Study (PHS)? The interesting history that led to this experiment is detailed at the PHS Web site. You can also find out about the Physicians' Health Study II, which ended in December 2007.

The control group in the Physicians' Health Study received inactive versions (placebos) of both aspirin and beta-carotene. This group served as a comparison for the effects of aspirin alone, beta-carotene alone, and aspirin plus beta-carotene.

## Experiments: What Can Go Wrong?

The logic of a randomized comparative experiment depends on our ability to treat all the subjects the same in every way except for the actual treatments being compared. Good experiments, therefore, require careful attention to details to ensure that all subjects really are treated identically.

If some subjects in a medical experiment take a pill each day and a control group takes no pill, the subjects are not treated identically. Many medical experiments are therefore "placebo-controlled," like the Physicians' Health Study. All the subjects received the same medical attention during the several years of the experiment. On odd-numbered days, all of them took an aspirin or a placebo. On even-numbered days, all subjects took either a beta-carotene pill or a placebo. Many patients respond favorably to any treatment, even a placebo, perhaps because they trust the doctor. The response to a dummy treatment is called the

## Placebo effect

**placebo effect.** If some subjects did not take any pills, the effect of aspirin or beta-carotene would be confounded with the placebo effect, the effect of simply taking pills.

**EXAMPLE***Curing Baldness and Soothing Pain***Do placebos work?**

Want to help balding men keep their hair? Give them a placebo—one study found that 42% of balding men maintained or increased the amount of hair on their heads when they took a placebo. In another study, researchers zapped the wrists of 24 test subjects with a painful jolt of electricity. Then they rubbed a cream with no active medicine on subjects' wrists and told them the cream should help soothe the pain. When researchers shocked them again, 8 subjects said they experienced significantly less pain.<sup>24</sup>

When the ailment is vague and psychological, like depression, some experts think that about three-quarters of the effect of the most widely used drugs is just the placebo effect.<sup>25</sup> Others disagree. The strength of the placebo effect in medical treatments is hard to pin down because it depends on the exact environment. How enthusiastic the doctor is seems to matter a lot. But “placebos work” is a good place to start when you think about planning medical experiments.

B.C.

by johnny hart



Reports in medical journals regularly begin with words like these, from a study of a flu vaccine given as a nose spray: “This study was a randomized, double-blind, placebo-controlled trial. Participants were enrolled from 13 sites across the continental United States between mid-September and mid-November.”<sup>26</sup> Doctors are supposed to know what this means. Now you know, too.

The strength of the placebo effect is a strong argument for randomized comparative experiments. In the baldness study, 42% of the placebo group kept or increased their hair, but 86% of the men getting a new drug to fight baldness did so. The drug beats the placebo, so it has something besides the placebo effect going for it. Of course, the placebo effect is still part of the reason this and other treatments work.

Because the placebo effect is so strong, it would be foolish to tell subjects in a medical experiment whether they are receiving a new drug or a placebo. Knowing that they are getting “just a placebo” might weaken the placebo effect and bias the experiment in favor of the other treatments. It is also foolish to tell doctors and other medical personnel what treatment each subject received. If they know that a subject is getting “just a placebo,” they may expect less than if they know the subject is receiving a promising experimental drug. Doctors’ expectations change how they interact with patients and even the way they diagnose a patient’s condition. Whenever possible, experiments with human subjects should be **double-blind**.

**DEFINITION: Double-blind**

In a **double-blind** experiment, neither the subjects nor those who interact with them and measure the response variable know which treatment a subject received.

## Single-blind

The idea of a double-blind design is simple. Until the experiment ends and the results are in, only the study's statistician knows for sure which treatment a subject is receiving. However, some experiments cannot be carried out in a double-blind manner. If researchers are comparing the effects of exercise and dieting on weight loss, then subjects will know which treatment they are receiving. Such an experiment can still be **single-blind** if the individuals who are interacting with the subjects and measuring the response variable don't know who is dieting and who is exercising. In other single-blind experiments, the subjects are unaware of which treatment they are receiving, but the people interacting with them and measuring the response variable do know.

**CHECK YOUR UNDERSTANDING**

In an interesting experiment, researchers examined the effect of ultrasound on birth weight. Pregnant women participating in the study were randomly assigned to one of two groups. The first group of women received an ultrasound; the second group did not. When the subjects' babies were born, their birth weights were recorded. The women who received the ultrasounds had heavier babies.<sup>27</sup>

1. Did the experimental design take the placebo effect into account? Why is this important?
2. Was the experiment double-blind? Why is this important?
3. Based on your answers to Questions 1 and 2, describe an improved design for this experiment.

## Inference for Experiments

In an experiment, researchers usually hope to see a difference in the responses so large that it is unlikely to happen just because of chance variation. We can use the laws of probability, which describe chance behavior, to learn whether the treatment effects are larger than we would expect to see if only chance were operating. If they are, we call them **statistically significant**.

**DEFINITION: Statistically significant**

An observed effect so large that it would rarely occur by chance is called **statistically significant**.

If we observe statistically significant differences among the groups in a randomized comparative experiment, we have good evidence that the treatments actually caused these differences. You will often see the phrase “statistically significant” in published research reports in many fields. The great advantage of randomized comparative experiments is that they can produce data that give good evidence for a cause-and-effect relationship between the explanatory and response variables. We know that in general a strong association does not imply causation. A statistically significant association in data from a well-designed experiment *does* imply causation.



## ACTIVITY *Distracted driving*

**MATERIALS:** Deck of cards for each team of 3 to 4 students



Is talking on a cell phone while driving more distracting than talking to a passenger? David Strayer and his colleagues at the University of Utah designed an experiment to help answer this question. They used 48 undergraduate students as subjects. The researchers randomly assigned half of the subjects to drive in a simulator while talking on a cell phone, and the other half to drive in the simulator while talking to a passenger. One response variable was whether the driver stopped at a rest area that was specified by researchers before the simulation started. The table below shows the results:<sup>28</sup>

Stopped at rest area?	Distraction	
	Cell phone	Passenger
Yes	12	21
No	12	3

Are these results statistically significant? To find out, let's see what would happen just by chance if we randomly reassign the 48 people in this experiment to the two groups many times, *assuming the treatment received doesn't affect whether a driver stops at the rest area*.

1. We need 48 cards from the deck to represent the drivers in this study. In the original experiment, 33 drivers stopped at the rest area and 15 didn't. Since we're assuming that the treatment received won't change whether each driver stops at the rest area, we use 33 cards to represent drivers who stop and 15 cards to represent those who don't. Remove the ace of spades and any three of the 2s from the deck.

- Stop: All cards with denominations 2 through 10 ( $36 - 3 \text{ missing } 2\text{s} = 33$ )
- Don't stop: All jacks, queens, kings, and aces ( $16 - 1 \text{ missing ace} = 15$ )

2. Shuffle and deal two piles of 24 cards each—the first pile represents the cell phone group and the second pile represents the passenger group. The shuffling reflects our assumption that the outcome for each subject is not affected by the treatment. Record the number of drivers who failed to stop at the rest area in each group.

3. Repeat this process 9 more times so that you have a total of 10 trials. Record your results in a table like this:

Trial	Number who didn't stop in cell phone group	Number who didn't stop in passenger group
1		

In how many of your 10 trials did 12 or more drivers in the cell phone group fail to stop?

4. Make a class dotplot of the number of drivers in the cell phone group who failed to stop at the rest area in each trial. In what percent of the class's trials did 12 or more people in the cell phone group fail to stop at the rest area?

5. In the original experiment, 12 of the 24 drivers using cell phones didn't stop at the rest area. Based on the class's simulation results, how surprising would it be to get a result this large or larger simply due to the chance involved in the random assignment? Is the result statistically significant?
6. What conclusion would you draw about whether talking on a cell phone is more distracting than talking to a passenger?

## THINK ABOUT IT

Can an “unlucky” random assignment lead to confounding? Let's return to the distracted-driver Activity. Some people are more forgetful than others. Suppose that the random assignment happens to put most of the forgetful subjects in one group. If more drivers in that group fail to stop at the rest area, we don't know if it's because of the treatment they received (cell phone or passenger) or their forgetfulness. Is this confounding?

You might be surprised that the answer is “No!” Although people's memory is a lurking variable that might affect whether they stop at the rest area (the response variable), the design of the experiment takes care of this by randomly assigning subjects to the two treatment groups. The “unlucky” random assignments are taken into account in determining statistical significance. In an experiment, confounding occurs when the design doesn't account for existing differences in the experimental units that might systematically affect their response to the treatments.

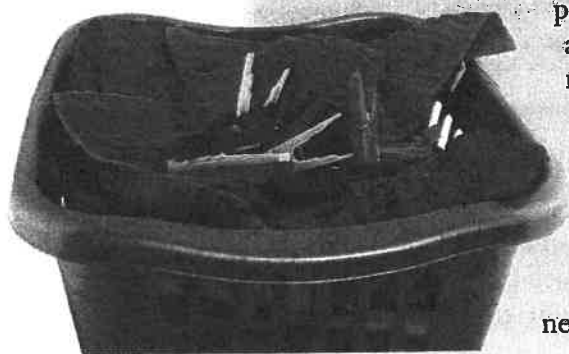
## Blocking

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of control, random assignment, and replication. But just as with sampling, there are times when the simplest method doesn't yield the most precise results. When a population consists of groups of individuals that are “similar within but different between,” a stratified random sample gives a better estimate than a simple random sample. This same logic applies in experiments.

### EXAMPLE

#### *Doing the Laundry* The idea of blocking

Suppose researchers want to test whether a new detergent for clothes that require hand-washing cleans better in warm or in cold water. They decide to perform an experiment using numerous pieces of dirty laundry as the experimental units. The response variable is a cleanliness rating on a scale of 0 (very dirty) to 10 (very clean). How should researchers deal with the fact that light-colored clothing tends to come cleaner in warm water? They could use a completely randomized design and hope that the random assignment distributes the light-colored clothing about evenly between the cold- and warm-water treatment groups. Even so, there might be a lot of variability in cleanliness ratings in both groups due to the color of the clothing. This

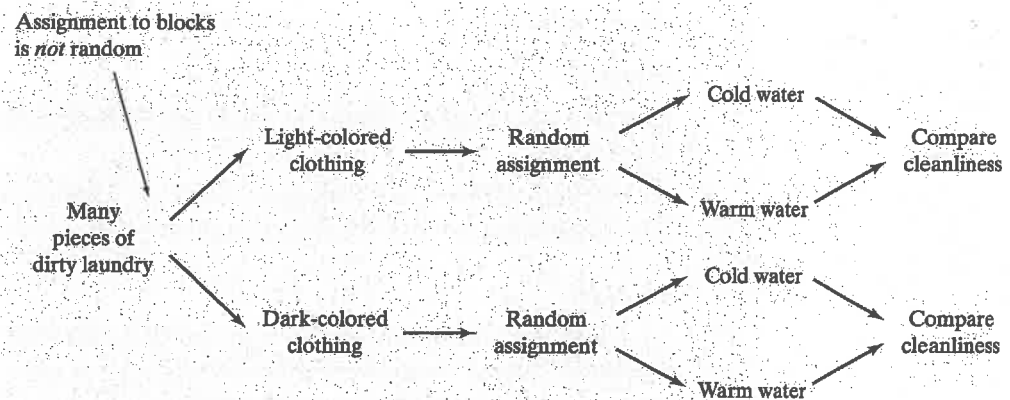


Color is a lurking variable in the completely randomized design. The random assignment prevents confounding due to this variable, however.

lurking variable might make it difficult to detect a difference in the effects of the treatments. What should the researchers do?

Since the researchers know that color may affect cleanliness in a systematic way, they could start by separating the laundry into two piles—one for lighter colors and one for darker colors. Each of these piles of similar experimental units is known as a **block**. Within each block, the researchers could then randomly assign about half the pieces of laundry to be hand-washed in cold water and the other half to be hand-washed in warm water. The same person should do all the washing as a form of control. This **randomized block design** reduces the effect of variation among the experimental units due to color on the response variable.

Figure 4.7 outlines the randomized block design for our laundry experiment. The experimental units are first separated into blocks to deal with the potential confounding variable, color. Then, the two treatments are randomly assigned within each block.



**FIGURE 4.7** Outline of a randomized block design for the laundry experiment. The blocks consist of light-colored and dark-colored pieces of laundry. The treatments are warm and cold water.

#### DEFINITION: Block and randomized block design

A **block** is a group of experimental units that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

In a **randomized block design**, the random assignment of experimental units to treatments is carried out separately within each block.

Blocks are another form of *control*. They control the effects of some outside variables by bringing those variables into the experiment to form the blocks. A randomized block design allows us to draw separate conclusions about each block, for example, about light-colored and dark-colored clothes in the laundry experiment. Blocking also allows more precise overall conclusions, because the

**AP EXAM TIP** Don't mix the language of experiments and the language of sample surveys or other observational studies. You will lose credit for saying things like "use a randomized block design to select the sample for this survey" or "this experiment suffers from nonresponse since some subjects dropped out during the study."

systematic differences between light- and dark-colored clothes can be removed when we study the overall effects of using this detergent in warm and cold water.

The idea of blocking is an important additional principle of experimental design. A wise experimenter will form blocks based on the most important unavoidable sources of variability (lurking variables) among the experimental units. Randomization will then average out the effects of the remaining lurking variables and allow an unbiased comparison of the treatments. The moral of the story is: *control what you can, block on what you can't control, and randomize to create comparable groups.*

## EXAMPLE

### Men, Women, and Advertising

#### Blocking in an experiment

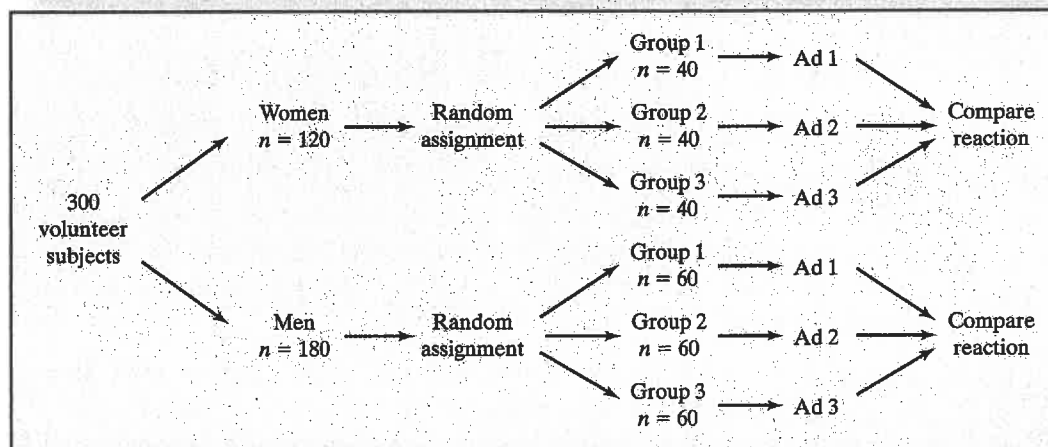
Women and men respond differently to advertising. Researchers would like to design an experiment to compare the effectiveness of three advertisements for the same product.

#### PROBLEM:

- Explain why a randomized block design might be preferable to a completely randomized design for this experiment.
- Outline a randomized block design using 300 volunteers (180 men and 120 women) as subjects. Describe how you would carry out the random assignment required by your design.

#### SOLUTION:

- A completely randomized design considers all subjects, both men and women, as a single pool. The random assignment would send subjects to three treatment groups without regard to their gender. This ignores the differences between men and women, which would probably result in a great deal of variability in responses to the advertising in all three groups. For example, if an ad appealed



**FIGURE 4.8** Randomized block design for comparing responses to three advertisements. The blocks consist of male and female subjects.

much more to men, you would get a wide range of reactions to that ad from the two genders. That would make it harder to compare the effectiveness of the ads.

A randomized block design would consider women and men separately. In this case, the random assignment would occur separately in each block. Blocking will control for the variability in responses to advertising due to gender. This will allow researchers to look separately at the reactions of men and women, as well as to more effectively assess the overall response to the ads.

(b) Figure 4.8 outlines the randomized block design. We randomly assign the 120 women into three groups of 40, one for each of the advertising treatments. Give each woman a distinct label between 1 and 120. Use a computer's random number generator to sort the numbers from 1 to 120 in a random order. The women with labels corresponding to the first 40 numbers in the list will view Ad 1, the next 40 will view Ad 2, and the last 40 will view Ad 3. Randomly assign the 180 men into three groups of 60 using a similar process. After each subject has viewed the assigned ad, compare reactions to the three ads within the gender blocks.

For Practice Try Exercise 79

## Matched Pairs Design

Matched pairs design

A common type of randomized block design for comparing two treatments is a **matched pairs design**. The idea is to create blocks by matching pairs of similar experimental units. Then we can use chance to decide which member of a pair gets the first treatment. The other subject in that pair gets the other treatment. That is, the random assignment of subjects to treatments is done within each matched pair. Just as with other forms of blocking, matching helps reduce the effect of variation among the experimental units.

Sometimes each "pair" in a matched pairs design consists of just one experimental unit that gets both treatments one after the other. In that case, each experimental unit serves as its own control. The *order* of the treatments can influence the response, so we randomize the order for each experimental unit.

### • ACTIVITY *Get your heart beating*

• MATERIALS: Clock or stopwatch

• Are standing pulse rates generally higher than sitting pulse rates? In this Activity, you will perform two experiments to try to answer this question.

1. *Completely randomized design* For the first experiment, you'll randomly assign half of the students in your class to stand and the other half to sit. You can use the hat method, Table D, or technology to carry out the random assignment. Once the two treatment groups have been formed, students should stand or sit as required. Then they should measure their pulses for one minute. Have the subjects in each group record their data on the board.

2. *Matched pairs design* In a matched pairs design, each student should receive both treatments in a random order. Since you already sat or stood in Step 1, you just need to do the opposite now. As before, everyone should measure their pulses for one minute after completing the treatment (that is, once they are standing or sitting). Have all the subjects record their data (both measurements) in a chart on the board.

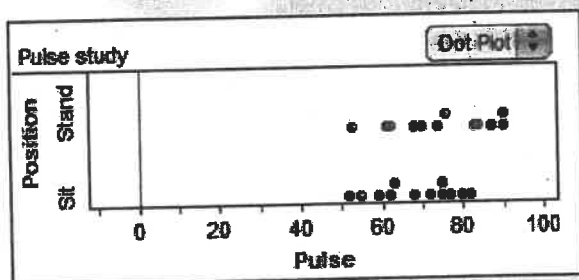


3. Analyze the data for the completely randomized design. Make a dotplot and calculate the mean pulse rate for each group. Is there evidence that standing pulse rates are higher? Explain.
4. Analyze the data for the matched pairs design. Since the data are paired by student, your first step should be to calculate the difference in pulse rate (standing – sitting) for each subject. Make a dotplot of these differences and calculate their mean. Is there evidence that standing pulse rates are higher? Explain.
5. What advantage does the matched pairs design have over the completely randomized design?

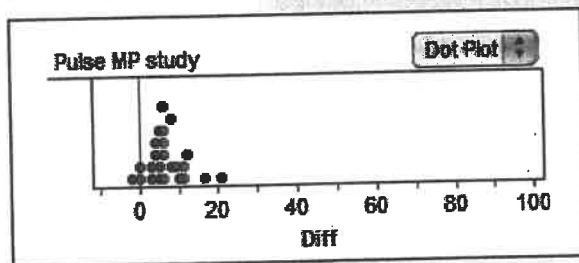
An AP Statistics class with 24 students performed the “Get Your Heart Beating” Activity. We’ll analyze the results of their experiment in the following example.

## EXAMPLE

### *Standing and Sitting Pulse Rate* Design determines analysis



A Fathom dotplot of the pulse rates for their completely randomized design is shown. The mean pulse rate for the standing group is 74.83; the mean for the sitting group is 68.33. So the average pulse rate is 6.5 beats per minute higher in the standing group. However, the variability in pulse rates for the two groups creates a lot of overlap in the graph. These data don’t provide convincing evidence that standing pulse rates tend to be higher.



What about the class’s matched pairs experiment? The Fathom dotplot shows their data on the difference in pulse rates (standing – sitting). For these 24 students, the mean difference was 6.8 beats per minute. In addition, 21 of the 24 students recorded a positive difference (meaning the standing pulse rate was higher). These data provide strong evidence that people’s standing pulse rates tend to be higher than their sitting pulse rates.

Let’s take one more look at the two Fathom dotplots in the example. Notice that we used the same scale for both graphs. This is to help you visually compare the amount of variability in the response variable for each of the two experimental designs. Blocking by subject in the matched pairs design greatly reduced the variability in the response variable. That made it easier to detect the fact that standing causes an increase in pulse rate. With the large amount of variability in the completely randomized design, we were unable to draw such a conclusion.

Another important lesson to take away from the example is this: *the design of the study determines the appropriate method of analysis*. For the completely randomized design, it makes sense to compare pulse rates for the two groups with parallel dotplots and means. In the matched pairs design, each student is a block. We compare the effects of the treatments within each block by examining the differences in standing and sitting pulse rates for each student.

The following Data Exploration asks you to apply what you have learned about analyzing data from an experiment.

### DATA EXPLORATION *Nitrogen in tires—a lot of hot air?*

Most automobile tires are inflated with compressed air, which consists of about 78% nitrogen. Aircraft tires are filled with pure nitrogen, which is safer than air in case of fire. Could filling automobile tires with nitrogen improve safety, performance, or both?

Consumers Union designed a study to test whether nitrogen-filled tires would maintain pressure better than air-filled tires. They obtained two tires from each of several brands and then filled one tire in each pair with air and one with nitrogen. All tires were inflated to a pressure of 30 pounds per square inch and then placed outside for a year. At the end of the year, Consumers Union measured the pressure in each tire. The amount of pressure lost (in pounds per square inch) during the year for the air-filled and nitrogen-filled tires of each brand is shown in the table below.<sup>29</sup>

Brand	Air	Nitrogen	Brand	Air	Nitrogen
BF Goodrich Traction T/A HR	7.6	7.2	Pirelli P6 Four Seasons	4.4	4.2
Bridgestone HP50 (Sears)	3.8	2.5	Sumitomo HTR H4	1.4	2.1
Bridgestone Potenza G009	3.7	1.6	Yokohama Avid H4S	4.3	3.0
Bridgestone Potenza RE950	4.7	1.5	BF Goodrich Traction T/A V	5.5	3.4
Bridgestone Potenza EL400	2.1	1.0	Bridgestone Potenza RE950	4.1	2.8
Continental Premier Contact H	4.9	3.1	Continental ContiExtreme Contact	5.0	3.4
Cooper Lifeliner Touring SLE	5.2	3.5	Continental ContiProContact	4.8	3.3
Dayton Daytona HR	3.4	3.2	Cooper Lifeliner Touring SLE	3.2	2.5
Falken Ziex ZE-512	4.1	3.3	General Exclaim UHP	6.8	2.7
Fuzion Hrl	2.7	2.2	Hankook Ventus V4 H105	3.1	1.4
General Exclaim	3.1	3.4	Michelin Energy MXV4 Plus	2.5	1.5
Goodyear Assurance TripleTred	3.8	3.2	Michelin Pilot Exalto A/S	6.6	2.2
Hankook Optimo H418	3.0	0.9	Michelin Pilot HX MXM4	2.2	2.0
Kumho Solus KH16	6.2	3.4	Pirelli P6 Four Seasons	2.5	2.7
Michelin Energy MXV4 Plus	2.0	1.8	Sumitomo HTR+	4.4	3.7
Michelin Pilot XGT H4	1.1	0.7			

Does filling tires with nitrogen instead of compressed air reduce pressure loss? Give appropriate graphical and numerical evidence to support your answer.

## SECTION 4.2

## Summary

- We can produce data intended to answer specific questions by **observational studies** or **experiments**. An observational study gathers data on individuals as they are. Experiments actively do something to people, animals, or objects in order to measure their response.
- Statistical studies often try to show that changing one variable (the **explanatory variable**) causes changes in another variable (the **response variable**). Variables are **confounded** when their effects on a response can't be distinguished from each other. Observational studies and uncontrolled experiments often fail to show that changes in an explanatory variable actually cause changes in a response variable because the explanatory variable is confounded with **lurking variables**.
- In an experiment, we impose one or more **treatments** on a group of **experimental units** (sometimes called **subjects** if they are human). Each treatment is a combination of values of the explanatory variables (also called **factors**).
- The **design** of an experiment describes the choice of treatments and the manner in which the subjects are assigned to the treatments. The basic principles of experimental design are **control** for lurking variables, **random assignment** of treatments, and **replication** (using enough experimental units).
  - The simplest form of *control* is comparison. Experiments should compare two or more treatments to avoid confounding of the effect of a treatment with other influences, such as lurking variables.
  - *Random assignment* uses chance to assign subjects to the treatments. Random assignment creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias**, or systematic favoritism, in experiments.
  - Applying each treatment to many experimental units (*replication*) reduces the role of chance variation and makes the experiment more likely to detect differences in the effects of the treatments.
- Many behavioral and medical experiments are **double-blind**. That is, neither the subjects nor those interacting with them and measuring their responses know who is receiving which treatment. If one party knows and the other doesn't, then the experiment is **single-blind**.
- Some experiments give a **placebo** (fake treatment) to a control group. That helps prevent confounding due to the **placebo effect**, in which some patients get better because they expect the treatment to work even though they have received an inactive treatment.
- In addition to comparison, a second form of control is to form **blocks** of individuals that are similar in some way that is important to the response. Randomization is then carried out separately within each block. Blocking helps reduce unwanted variability among experimental units.
- **Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, the subjects are matched in pairs as closely as possible, and each subject in a pair receives one of the treatments.

## SECTION 4.2

## Exercises

45. **Chocolate and happy babies** A University of Helsinki (Finland) study wanted to determine if chocolate consumption during pregnancy had an effect on infant temperament at age 6 months. Researchers began by asking 305 healthy pregnant women to report their chocolate consumption. Six months after birth, the researchers asked mothers to rate their infants' temperament, including smiling, laughter, and fear. The babies born to women who had been eating chocolate daily during pregnancy were found to be more active and "positively reactive"—a measure that the investigators said encompasses traits like smiling and laughter.<sup>30</sup>
- Was this an observational study or an experiment? Justify your answer.
  - What are the explanatory and response variables?
  - Does this study show that eating chocolate regularly during pregnancy helps produce infants with good temperament? Explain.
46. **Child care and aggression** A study of child care enrolled 1364 infants and followed them through their sixth year in school. Later, the researchers published an article in which they stated that "the more time children spent in child care from birth to age four-and-a-half, the more adults tended to rate them, both at age four-and-a-half and at kindergarten, as less likely to get along with others, as more assertive, as disobedient, and as aggressive."<sup>31</sup>
- Is this an observational study or an experiment? Justify your answer.
  - What are the explanatory and response variables?
  - Does this study show that child care causes children to be more aggressive? Explain.
47. **Learning biology with computers** An educator wants to compare the effectiveness of computer software for teaching biology with that of a textbook presentation. She gives a biology pretest to each of a group of high school juniors, then randomly divides them into two groups. One group uses the computer, and the other studies the text. At the end of the year, she tests all the students again and compares the increase in biology test scores in the two groups.
- Is this an observational study or an experiment? Justify your answer.
  - If the group using the computer has a much higher average increase in test scores than the group using the textbook, what conclusions, if any, could the educator draw?
48. **Cell phones and brain cancer** One study of cell phones and the risk of brain cancer looked at a group of 469 people who have brain cancer. The investigators matched each cancer patient with a person of the same age, gender, and race who did not have brain cancer, then asked about the use of cell phones. Result: "Our data suggest that the use of handheld cellular phones is not associated with risk of brain cancer."<sup>32</sup>
- Is this an observational study or an experiment? Justify your answer.
  - Based on this study, would you conclude that cell phones do not increase the risk of brain cancer? Why or why not?
49. **Effects of class size** Do smaller classes in elementary school really benefit students in areas such as scores on standardized tests, staying in school, and going on to college? We might do an observational study that compares students who happened to be in smaller and larger classes in their early school years. Identify a lurking variable that may lead to confounding with the effects of small classes. Explain how confounding might occur.
50. **Effects of binge drinking** A common definition of "binge drinking" is 5 or more drinks at one sitting for men and 4 or more for women. An observational study finds that students who binge drink have lower average GPA than those who don't. Identify a lurking variable that may be confounded with the effects of binge drinking. Explain how confounding might occur.
- For the experiments described in Exercises 51 to 56, identify the experimental units or subjects, the explanatory variables (factors), the treatments, and the response variables.*
51. **Growing in the shade** Ability to grow in shade may help pines found in the dry forests of Arizona to resist drought. How well do these pines grow in shade? Investigators planted pine seedlings in a greenhouse in either full light, light reduced to 25% of normal by shade cloth, or light reduced to 5% of normal. At the end of the study, they dried the young trees and weighed them.
52. **Internet telephone calls** You can use Voice over Internet Protocol (VoIP) to make long-distance telephone calls over the Internet. How will the cost affect the use of this service? A university plans an experiment to find out. It will offer the service to all

350 students in one of its dormitories. Some students will pay a low flat rate. Others will pay higher rates at peak periods and very low rates off-peak. The university is interested in the amount and time of use and in the effect on the congestion of the network.

53. **Improving response rate** How can we reduce the rate of refusals in telephone surveys? Most people who answer at all listen to the interviewer's introductory remarks and then decide whether to continue. One study made telephone calls to randomly selected households to ask opinions about the next election. In some calls, the interviewer gave her name, in others she identified the university she was representing, and in still others she identified both herself and the university. For each type of call, the interviewer either did or did not offer to send a copy of the final survey results to the person interviewed. Do these differences in the introduction affect whether the interview is completed?

54. **Eat well and exercise** Most American adolescents don't eat well and don't exercise enough. Can middle schools increase physical activity among their students? Can they persuade students to eat better? Investigators designed a "physical activity intervention" to increase activity in physical education classes and during leisure periods throughout the school day. They also designed a "nutrition intervention" that improved school lunches and offered ideas for healthy home-packed lunches. Each participating school was randomly assigned to one of the interventions, both interventions, or no intervention. The investigators observed physical activity and lunchtime consumption of fat.

- pg 235 ... 55. **Fabric science** A maker of fabric for clothing is setting up a new line to "finish" the raw fabric. The line will use either metal rollers or natural-bristle rollers to raise the surface of the fabric; a dyeing-cycle time of either 30 or 40 minutes; and a temperature of either 150° or 175° Celsius. An experiment will compare all combinations of these choices. Three specimens of fabric will be subjected to each treatment and scored for quality.

56. **Exercise and heart rate** A student project measured the increase in the heart rates of fellow students when they stepped up and down for three minutes to the beat of a metronome. The step was either 5.75 or 11.5 inches high and the metronome beat was 14, 21, or 28 steps per minute. Five students stepped at each combination of height and speed.

57. **Cocoa and blood flow** A study conducted by Norman Hollenberg, professor of medicine at Brigham and Women's Hospital and Harvard Medical School, involved 27 healthy people aged 18 to 72. Each

subject consumed a cocoa beverage containing 900 milligrams of flavonols (a class of flavonoids) daily for five days. Using a finger cuff, blood flow was measured on the first and fifth days of the study. After five days, researchers measured what they called "significant improvement" in blood flow and the function of the cells that line the blood vessels.<sup>33</sup> What flaw in the design of this experiment makes it impossible to say whether the cocoa really caused the improved blood flow? Explain.

58. **Reducing unemployment** Will cash bonuses speed the return to work of unemployed people? A state department of labor notes that last year 68% of people who filed claims for unemployment insurance found a new job within 15 weeks. As an experiment, this year the state offers \$500 to people filing unemployment claims if they find a job within 15 weeks. The percent who do so increases to 77%. What flaw in the design of this experiment makes it impossible to say whether the bonus really caused the increase? Explain.

- pg 237 ... 59. **Layoffs and "survivor guilt"** Workers who survive a layoff of other employees at their location may suffer from "survivor guilt." A study of survivor guilt and its effects used as subjects 120 students who were offered an opportunity to earn extra course credit by doing proofreading. Each subject worked in the same cubicle as another student, who was an accomplice of the experimenters. At a break midway through the work, one of three things happened:

*Treatment 1:* The accomplice was told to leave; it was explained that this was because she performed poorly.  
*Treatment 2:* It was explained that unforeseen circumstances meant there was only enough work for one person. By "chance," the accomplice was chosen to be laid off.

*Treatment 3:* Both students continued to work after the break.

The subjects' work performance after the break was compared with performance before the break.<sup>34</sup> Describe how you would randomly assign the subjects to the treatments

- (a) using slips of paper.
- (b) using Table D.
- (c) using technology.

60. **Effects of TV advertising** Figure 4.3 (page 235) displays the 6 treatments for a two-factor experiment on TV advertising. Suppose we have 150 students who are willing to serve as subjects. Describe how you would randomly assign the subjects to the treatments

- (a) using slips of paper.
- (b) using Table D.
- (c) using technology.



pg 239

61. **Headache relief** Doctors identify “chronic tension-type headaches” as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone? Investigators compared four treatments: antidepressant alone, placebo alone, antidepressant plus stress management, and placebo plus stress management. The headache sufferers named in the following table have agreed to participate in the study.

Acosta	Duncan	Han	Liang	Padilla	Valasco
Asihiro	Durr	Howard	Maldonado	Plochman	Vaughn
Bennett	Edwards	Hruska	Marsden	Rosen	Wei
Bikalis	Farouk	Imrani	Montoya	Solomon	Wilder
Chen	Fleming	James	O'Brian	Trujillo	Willis
Clemente	George	Kaplan	Ogle	Tullock	Zhang

- (a) Outline the design of the experiment. What is this type of design called?
- (b) Explain how you would randomly assign the subjects to the four treatment groups. Then carry out your random assignment.
62. **More rain for California?** The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers carried out an experiment to study the effects of more rain in either season. They randomly assigned plots of open grassland to 3 treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Thirty-six circular plots of area 70 square meters were available (see the photo), of which 18 were used for this study. One response variable was total plant biomass, in grams per square meter, produced in a plot over a year.<sup>35</sup>



- (a) Outline the design of the experiment. What is this type of design called?
- (b) Explain how you would randomly assign the experimental units to the three treatments. Then carry out your random assignment.

63. **Treating prostate disease** A large study used records from Canada's national health care system to compare the effectiveness of two ways to treat prostate disease. The two treatments are traditional surgery and a new method that does not require surgery. The records described many patients whose doctors had chosen each method. The study found that patients treated by the new method were significantly more likely to die within 8 years.<sup>36</sup>

(a) Further study of the data showed that this conclusion was wrong. The extra deaths among patients who got the new method could be explained by lurking variables. What lurking variables might be confounded with a doctor's choice of surgical or nonsurgical treatment?

(b) You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Write a few sentences describing how you would design this experiment.

64. **Getting teachers to come to school** Elementary schools in rural India are usually small, with a single teacher. The teachers often fail to show up for work. Here is an idea for improving attendance: give the teacher a digital camera with a tamperproof time and date stamp and ask a student to take a photo of the teacher and class at the beginning and end of the day. Offer the teacher better pay for good attendance, verified by the photos. Will this work? Researchers obtained permission to use 120 rural schools in Rajasthan for an experiment to find out.<sup>37</sup>

(a) Explain why it would not be a good idea to offer better pay for good attendance to the teachers in all 120 schools and then to compare this year's attendance with last year's.

(b) Write a few sentences describing how you would design this experiment.

65. **Stronger players** A football coach hears that a new exercise program will increase upper-body strength better than lifting weights. He is eager to test this new program in the off-season with the players on his high school team. The coach decides to let his players choose which of the two treatments they will undergo for 3 weeks—exercise or weight lifting. He will use the number of push-ups a player can do at the end of the experiment as the response variable.

(a) Which principle of experimental design does the coach's plan violate? Explain how this violation could lead to confounding.

(b) Comment on the coach's choice of response variable.

66. **Prayer and meditation** You read in a magazine that “nonphysical treatments such as meditation and prayer have been shown to be effective in controlled scientific studies for such ailments as high blood

pressure, insomnia, ulcers, and asthma." Explain in simple language what the article means by "controlled scientific studies." Why can such studies provide good evidence that meditation is an effective treatment for high blood pressure?

pg 241

67. **The effects of day care** Does day care help low-income children stay in school and hold good jobs later in life? The Carolina Abecedarian Project (the name suggests the ABCs) has followed a group of 111 children since 1972. Back then, these individuals were all healthy but low-income black infants in Chapel Hill, North Carolina. All the infants received nutritional supplements and help from social workers. Half were also assigned at random to an intensive preschool program.<sup>38</sup>

- Explain the purpose of each of the three experimental design principles.
- Describe how each of these principles was used in this study.

68. **Killing weeds** A biologist would like to determine which of two brands of weed killer is less likely to harm the plants in a garden at the university. Before spraying near the plants, the biologist decides to conduct an experiment using 24 individual plants. Which of the following two plans for randomly assigning the treatments should the biologist use? Why?

*Plan A:* Choose the 12 healthiest-looking plants. Apply Brand X weed killer to all 12 of those plants. Apply Brand Y weed killer to the remaining 12 plants.

*Plan B:* Choose 12 of the 24 plants at random. Apply Brand X weed killer to those 12 plants and Brand Y weed killer to the remaining 12 plants.

69. **Do placebos really work?** Researchers in Japan conducted an experiment on 13 individuals who were extremely allergic to poison ivy. On one arm, each subject was rubbed with a poison ivy leaf and told the leaf was harmless. On the other arm, each subject was rubbed with a harmless leaf and told it was poison ivy. All the subjects developed a rash on the arm where the harmless leaf was rubbed. Of the 13 subjects, 11 did not have any reaction to the real poison ivy leaf.<sup>39</sup>

- What was the placebo in this experiment?
- Explain how the results of this study support the idea of a placebo effect.

70. **Pain relief study** Fizz Laboratories, a pharmaceutical company, has developed a new drug for relieving chronic pain. Sixty patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, "About what percent of pain relief did you experience?"

- Why should Fizz not simply administer the new drug and record the patients' responses?
- Should the patients be told whether they are getting the new drug or a placebo? How would this knowledge probably affect their reactions?

71. **Meditation for anxiety** An experiment that claimed to show that meditation lowers anxiety proceeded as follows. The experimenter interviewed the subjects and rated their level of anxiety. Then the subjects were randomly assigned to two groups. The experimenter taught one group how to meditate and they meditated daily for a month. The other group was simply told to relax more. At the end of the month, the experimenter interviewed all the subjects again and rated their anxiety level. The meditation group now had less anxiety. Psychologists said that the results were suspect because the ratings were not blind. Explain what this means and how lack of blindness could bias the reported results.

72. **Testosterone for older men** As men age, their testosterone levels gradually decrease. This may cause a reduction in lean body mass, an increase in fat, and other undesirable changes. Do testosterone supplements reverse some of these effects? A study in the Netherlands assigned 237 men aged 60 to 80 with low or low-normal testosterone levels to either a testosterone supplement or a placebo. The report in the *Journal of the American Medical Association* described the study as a "double-blind, randomized, placebo-controlled trial."<sup>40</sup> Explain each of these terms to someone who knows no statistics.

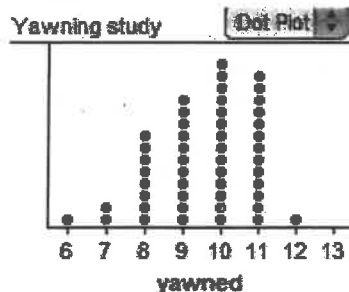
73. **Acupuncture and pregnancy** A study sought to determine whether the ancient Chinese art of acupuncture could help infertile women become pregnant.<sup>41</sup> One hundred sixty healthy women undergoing treatment with artificial insemination were recruited for the study. Half of the subjects were randomly assigned to receive acupuncture treatment 25 minutes before embryo transfer and again 25 minutes after the transfer. The remaining 80 subjects were instructed to lie still for 25 minutes after the embryo transfer. *Results:* In the acupuncture group, 34 women became pregnant. In the control group, 21 women became pregnant.

- Describe how the three principles of experimental design were addressed in this study.
- The difference in the percent of women who became pregnant in the two groups is statistically significant. Explain what this means to someone who knows little statistics.
- Explain why the placebo effect prevents us from concluding that acupuncture caused the difference in pregnancy rates.

74. **Do reducing diets work?** Dr. Linda Stern and her colleagues recruited 132 obese adults at the Philadelphia Veterans Affairs Medical Center in Pennsylvania. Half the participants were randomly assigned to a low-carbohydrate diet and the other half to a low-fat diet. Researchers measured each participant's change in weight and cholesterol level after six months and again after one year. Subjects in the low-carb diet group lost significantly more weight than subjects in the low-fat diet group during the first six months. At the end of a year, however, the average weight loss for subjects in the two groups was not significantly different.<sup>42</sup>

- Why did researchers randomly assign the subjects to the diet treatments?
  - Explain to someone who knows little statistics what "lost significantly more weight" means.
  - The subjects in the low-carb diet group lost an average of 5.1 kg in a year. The subjects in the low-fat diet group lost an average of 3.1 kg. Explain how this information could be consistent with the fact that weight loss in the two groups was not significantly different.
75. **Is yawning contagious?** According to the popular TV show *Mythbusters*, the answer is "Yes." In the March 9, 2005, episode, the *Mythbusters* team presented the results of an experiment involving 50 subjects. All the subjects were placed in a booth for an extended period of time and monitored by hidden camera. Two-thirds of the subjects were given a "yawn seed" by one of the experimenters; that is, the experimenter yawned in the subject's presence prior to leaving the room. The remaining subjects were given no yawn seed. What were the results? Of the 16 subjects who had no yawn seed, 4 yawned. Of the 34 subjects given a yawn seed, 10 yawned. Adam Savage and Jamie Hyneman, the cohosts of *Mythbusters*, used these results to conclude that yawning is contagious.

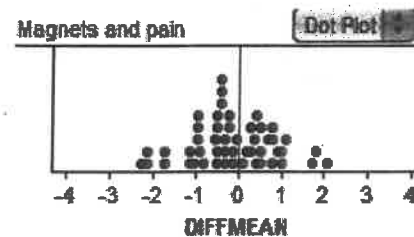
- Explain how you could use slips of paper to randomly *reassign* the subjects to the treatment groups, assuming that the treatment received doesn't affect the response.
- Suppose we used your method in (a) to redo the random assignment 50 times. The Fathom dotplot displays the number of subjects in the yawn seed group who yawned in each of these random assignments. What conclusion would you draw about whether yawning is contagious? Explain.



76. **Magnets and pain** Refer to the chapter-opening Case Study (page 205). The researchers decided to analyze the patients' final pain ratings. It also makes sense to examine the *difference* between patients' initial pain ratings and their final pain ratings in both groups. Here are the data:

**Active:** 10, 6, 1, 10, 6, 8, 5, 5, 6, 8, 7, 8, 7, 6, 4, 4, 7, 10, 6, 10, 6, 5, 5, 1, 0, 0, 0, 0, 1  
**Inactive:** 4, 3, 5, 2, 1, 4, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1

- Construct a comparative dotplot of the data. Describe what you see.
- Calculate the mean (average) change in pain rating for each group. Find the difference in the average changes for the two groups (Active – Inactive).
- Describe how you could use index cards to randomly reassign the subjects to the treatment groups, assuming that the treatment received doesn't affect the response.
- Suppose we used your method in (c) to redo the random assignment 50 times. The Fathom dotplot displays the difference (Active – Inactive) for the average change in pain rating for each of these random assignments. What conclusion would you draw about the effect of magnets on pain relief? Explain.

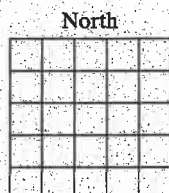


77. **Doctors and nurses** Nurse-practitioners are nurses with advanced qualifications who often act much like primary-care physicians. Are they as effective as doctors at treating patients with chronic conditions? An experiment was conducted with 1316 patients who had been diagnosed with asthma, diabetes, or high blood pressure. Within each condition, patients were randomly assigned to either a doctor or a nurse-practitioner. The response variables included measures of the patients' health and of their satisfaction with their medical care after 6 months.<sup>43</sup>
- Which are the blocks in this experiment: the different diagnoses (asthma, etc.) or the type of care (nurse or doctor)? Why?
  - Explain why a randomized block design is preferable to a completely randomized design here.
78. **Comparing cancer treatments** The progress of a type of cancer differs in women and men. Researchers want to design an experiment to compare three therapies for this cancer. They recruit 500 male and 300 female patients who are willing to serve as subjects.

- (a) Which are the blocks in this experiment: the cancer therapies or the two sexes? Why?
- (b) What are the advantages of a randomized block design over a completely randomized design using these 800 subjects?
- (c) Suppose the researchers had 800 male and no female subjects available for the study. What advantage would this offer? What disadvantage?

pg 248

79. **In the cornfield** An agriculture researcher wants to compare the yield of 5 corn varieties: A, B, C, D, and E. The field in which the experiment will be carried out increases in fertility from north to south. The researcher therefore divides the field into 25 plots of equal size, arranged in 5 east-west rows of 5 plots each, as shown in the diagram.



- (a) Explain why a randomized block design would be better than a completely randomized design in this setting.
- (b) Should the researcher use the rows or the columns of the field as blocks? Justify your answer.
- (c) Use technology or Table D to carry out the random assignment required by your design. Explain your method clearly.
80. **Comparing weight-loss treatments** Twenty overweight females have agreed to participate in a study of the effectiveness of four weight-loss treatments: A, B, C, and D. The researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are as follows:

Birnbaum	35	Hernandez	25	Moses	25	Smith	29
Brown	34	Jackson	33	Nevesky	39	Stall	33
Brunk	30	Kendall	28	Obrach	30	Tran	35
Cruz	34	Loren	32	Rodriguez	30	Wilansky	42
Deng	24	Mann	28	Santiago	27	Williams	22

The response variable is the weight lost after 8 weeks of treatment. Previous studies have shown that the effects of a diet may vary based on a subject's initial weight.

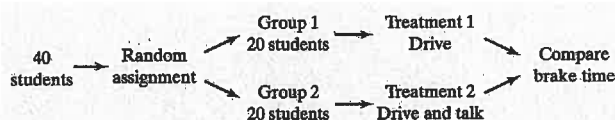
- (a) Explain why a randomized block design would be better than a completely randomized design in this setting.
- (b) Should researchers form blocks of size 4 based on subjects' last names in alphabetical order or by how overweight the subjects are? Explain.

- (c) Use technology or Table D to carry out the random assignment required by your design. Explain your method clearly.

81. **Aw, rats!** A nutrition experimenter intends to compare the weight gain of newly weaned male rats fed Diet A with that of rats fed Diet B. To do this, she will feed each diet to 10 rats. She has available 10 rats from one litter and 10 rats from a second litter. Rats in the first litter appear to be slightly healthier.
- (a) If the 10 rats from Litter 1 were fed Diet A, the effects of genetics and diet would be confounded, and the experiment would be biased. Explain this statement carefully.
- (b) Describe a better design for this experiment.
82. **Technology for teaching statistics** The Brigham Young University (BYU) statistics department is performing experiments to compare teaching methods. Response variables include students' final-exam scores and a measure of their attitude toward statistics. One study compares two levels of technology for large lectures: standard (overhead projectors and chalk) and multimedia. There are 8 lecture sections of a basic statistics course at BYU, each with about 200 students. There are four instructors, each of whom teaches two sections.<sup>44</sup> Suppose the sections and lecturers are as follows:

Section	Lecturer	Section	Lecturer
1	Hilton	5	Tolley
2	Christensen	6	Hilton
3	Hadfield	7	Tolley
4	Hadfield	8	Christensen

- (a) Suppose we randomly assign two lecturers to use standard technology in their sections and the other two lecturers to use multimedia technology. Explain how this could lead to confounding.
- (b) Describe a better design for this experiment.
83. **Look, Ma, no hands!** Does talking on a hands-free cell phone distract drivers? Researchers recruit 40 student subjects for an experiment to investigate this question. They have a driving simulator equipped with a hands-free phone for use in the study.
- (a) Researchers are considering the design shown in the figure below. What type of design is this?



- (b) Explain how blocking could be used to improve the design in (a).
- (c) Why is it important to randomly assign the treatments within each block?
84. **Chocolate gets my heart pumping** Cardiologists at Athens Medical School in Greece wanted to test whether chocolate affected blood flow in the blood vessels. The researchers recruited 17 healthy young volunteers, who were each given a 3.5-ounce bar of dark chocolate, either bittersweet or fake chocolate. On another day, the volunteers were switched. The subjects had no chocolate outside the study, and investigators didn't know whether a subject had eaten the real or the fake chocolate. An ultrasound was taken of each volunteer's upper arm to see the functioning of the cells in the walls of the main artery. The investigators found that blood vessel function was improved when the subjects ate bittersweet chocolate, and that there were no such changes when they ate the placebo (fake chocolate).<sup>45</sup>
- (a) What type of design did the investigators use in their study?
- (b) Explain why the investigators chose this design instead of a completely randomized design.
- (c) Why is it important to randomly assign the order of the treatments for the subjects?
85. **Room temperature and dexterity** An expert on worker performance is interested in the effect of room temperature on the performance of tasks requiring manual dexterity. She chooses temperatures of 70°F and 90°F as treatments. The response variable is the number of correct insertions, during a 30-minute period, in a peg-and-hole apparatus that requires the use of both hands simultaneously. Each subject is trained on the apparatus and then asked to make as many insertions as possible in 30 minutes of continuous effort.
- (a) Outline a completely randomized design to compare dexterity at 70° and 90°. Twenty subjects are available.
- (b) Because individuals differ greatly in dexterity, the wide variation in individual scores may hide the systematic effect of temperature unless there are many subjects in each group. Describe in detail the design of a matched pairs experiment in which each subject serves as his or her own control.
86. **Carbon dioxide and tree growth** The concentration of carbon dioxide (CO<sub>2</sub>) in the atmosphere is increasing rapidly due to our use of fossil fuels. Because plants use CO<sub>2</sub> to fuel photosynthesis, more CO<sub>2</sub> may cause trees and other plants to grow faster. An elaborate apparatus allows researchers to pipe extra CO<sub>2</sub> to a 30-meter circle of forest. We want to compare the growth in base area of trees in treated and untreated areas to see if extra CO<sub>2</sub> does in fact increase growth. We can afford to treat three circular areas.<sup>46</sup>
- (a) Describe the design of a completely randomized experiment using 6 well-separated 30-meter circular areas in a pine forest. Sketch the circles and carry out the randomization your design calls for.
- (b) Areas within the forest may differ in soil fertility. Describe a matched pairs design using three pairs of circles that will reduce the extra variation due to different fertility. Sketch the circles and carry out the randomization your design calls for.
87. **Got deodorant?** A group of students wants to perform an experiment to determine whether Brand A or Brand B deodorant lasts longer. One group member suggests the following design: Recruit 40 student volunteers—20 male and 20 female. Separate by gender, since male and female bodies might respond differently to deodorant. Give all the males Brand A deodorant and all the females Brand B. Have each student rate how well the deodorant is still working at the end of the school day on a 0 to 10 scale. Then compare ratings for the two treatments.
- (a) Identify any flaws you see in the proposed design for this experiment.
- (b) Describe how you would design the experiment. Explain how your design addresses each of the problems you identified in (a).
88. **Wake-up call** Do people naturally wake up earlier when they set an alarm before going to sleep? Justin decides to conduct his own experiment to find out. On Friday and Saturday nights, he doesn't set the alarm before going to bed. On Monday and Tuesday, he sets the alarm for 7 A.M. Justin records the time when he wakes up each day and then compares his average wake-up time with and without the alarm.
- (a) Identify any flaws you see in the proposed design for this experiment.
- (b) Describe how you would design the experiment. Explain how your design addresses each of the problems you identified in (a).
89. **Close shave** Which of two brands of electric razor shaves closer? Describe how you would design and carry out an experiment to answer this question using 50 adult male volunteers.
90. **Music and studying** Does listening to music while reading a story help or hinder recall of factual details? Describe how you would design and carry out an experiment to answer this question using 30 students at your school who have agreed to participate.



**Multiple choice: Select the best answer for Exercises 91 to 98.**

91. Can changing diet reduce high blood pressure? Vegetarian diets and low-salt diets are both promising. Men with high blood pressure are assigned at random to four diets: (1) normal diet with unrestricted salt; (2) vegetarian with unrestricted salt; (3) normal with restricted salt; and (4) vegetarian with restricted salt. This experiment has
- one factor, the type of diet.
  - two factors, high blood pressure and type of diet.
  - two factors, normal/vegetarian diet and unrestricted/restricted salt.
  - three factors, men, high blood pressure, and type of diet.
  - four factors, the four diets being compared.
92. In the experiment of the previous exercise, the 240 subjects are labeled 001 to 240. Software randomly assigns 60 subjects to each of the four diets. This is
- a completely randomized design.
  - a randomized block design.
  - a matched pairs design.
  - an observational study.
  - an SRS.
93. The Community Intervention Trial for Smoking Cessation asked whether a community-wide advertising campaign would reduce smoking. The researchers located 11 pairs of communities, each pair similar in location, size, economic status, and so on. One community in each pair participated in the advertising campaign and the other did not. This is
- an observational study.
  - a matched pairs experiment.
  - a completely randomized experiment.
  - a randomized block design, but not matched pairs.
  - a stratified random sample.
94. The drug manufacturer Merck recently stopped testing a promising new drug to treat depression. It turned out that in a randomized, double-blind trial, a dummy pill did almost as well as the new drug. The fact that many people respond to a dummy treatment is called
- confounding.
  - nonresponse.
  - comparison.
  - the placebo effect.
  - bias.
95. Corn variety 1 yielded 140 bushels per acre last year at a research farm. This year, corn variety 2, planted in the same location, yielded only 110 bushels per acre. Unfortunately, we don't know whether the difference is due to the superiority of variety 1 or to the effect of this year's drought. This is an example of
- bias.
  - matched pairs design.
  - confounding.
  - the placebo effect.
  - replication.
96. A report in a medical journal notes that the risk of developing Alzheimer's disease among subjects who (voluntarily) regularly took the anti-inflammatory drug ibuprofen (the active ingredient in Advil) was about half the risk among those who did not. Is this good evidence that ibuprofen is effective in preventing Alzheimer's disease?
- Yes, because the study was a randomized, comparative experiment.
  - No, because the effect of ibuprofen is confounded with the placebo effect.
  - Yes, because the results were published in a reputable professional journal.
  - No, because this is an observational study. An experiment would be needed to confirm (or not confirm) the observed effect.
  - Yes, because a 50% reduction can't happen just by chance.
97. A double-blind experiment was conducted to evaluate the effectiveness of the Salk polio vaccine. The purpose of keeping the diagnosing physicians unaware of the treatment status of the experimental subjects was to
- eliminate grounds for malpractice suits.
  - ensure that subjects were randomly assigned to treatments.
  - eliminate a possible source of bias.
  - make sure nobody is harmed.
  - avoid the placebo effect.
98. Two essential features of all statistically designed experiments are
- compare several treatments; use the double-blind method.
  - compare several treatments; use chance to assign subjects to treatments.
  - always have a placebo group; use the double-blind method.
  - use a block design; use chance to assign subjects to treatments.
  - use enough subjects; always have a control group.

99. **Seed weights (2.2)** Biological measurements on the same species often follow a Normal distribution quite closely. The weights of seeds of a variety of winged bean are approximately Normal with mean 525 milligrams (mg) and standard deviation 110 mg.
- (a) What percent of seeds weigh more than 500 mg? Show your method.
- (b) If we discard the lightest 10% of these seeds, what is the smallest weight among the remaining seeds? Show your method.

*Exercises 100 and 101 refer to the following setting. A researcher studied a group of identical twins who had been separated and adopted at birth. In each case, one twin (Twin A) was adopted by a low-income family and the*

*other (Twin B) by a high-income family. Both twins were given an IQ test as adults. Here are their scores:*<sup>47</sup>

Twin A:	120	99	99	94	111	97	99	94	104	114	113	100
Twin B:	128	104	108	100	116	105	100	100	103	124	114	112

100. **IQ and twins (3.1)** How well does one twin's IQ predict the other's? Give appropriate evidence to support your answer.
101. **IQ and income (1.3)** Do identical twins living in low-income homes tend to have lower IQs later in life than their twins who live in high-income homes? Give appropriate evidence to support your answer.

## 4.3

## Using Studies Wisely

**In Section 4.3, you'll learn about:**

- Scope of inference
- The challenges of establishing causation
- Data ethics

Researchers who conduct statistical studies often want to draw conclusions (make inferences) that go beyond the data they produce. Here are two examples.

- The U.S. Census Bureau carries out a monthly Current Population Survey of about 60,000 households. Their goal is to use data from these randomly selected households to estimate the percent of unemployed individuals in the population.
- Scientists performed an experiment that randomly assigned 21 volunteer subjects to one of two treatments: sleep deprivation for one night or unrestricted sleep. The experimenters hoped to show that sleep deprivation causes a decrease in performance two days later.<sup>48</sup>

What type of inference can be made from a particular study? The answer depends on the design of the study.

### Scope of Inference

In the Census Bureau's sample survey, the individuals who responded were chosen at random from the population of interest. Random sampling avoids bias and produces trustworthy estimates of the truth about the population. The Census Bureau should be safe making an *inference about the population* based on the results of the sample.

In the sleep deprivation experiment, subjects were randomly assigned to the sleep deprivation and unrestricted sleep treatments. Random assignment helps ensure that the two groups of subjects are as similar as possible before the treatments are imposed. If the unrestricted sleep group performs much better than the sleep deprivation group, and the difference is too large to be explained by chance variation in the random assignment, it must be due to the treatments. In that case, the scientists could safely conclude that sleep deprivation caused the decrease in performance. That is, they can make an *inference about cause and effect*. However, since the experiment used volunteer subjects, this limits scientists' ability to generalize their findings to some larger population of individuals.



Let's recap what we've learned about the scope of inference in a statistical study. Random selection of individuals allows inference about the population. Random assignment of individuals to groups permits inference about cause and effect. The chart below summarizes the possibilities.<sup>49</sup>

Both random sampling and random assignment introduce chance variation into a statistical study. When performing inference, statisticians use the laws of probability to describe this chance variation. You'll learn how this works later in the book.

		Were individuals randomly assigned to groups?	
		Yes	No
Were individuals randomly selected?	Yes	Inference about the population: YES Inference about cause and effect: YES	Inference about the population: YES Inference about cause and effect: NO
	No	Inference about the population: NO Inference about cause and effect: YES	Inference about the population: NO Inference about cause and effect: NO

Well-designed experiments randomly assign individuals to treatment groups. However, most experiments don't select experimental units at random from the larger population. That limits such experiments to inference about cause and effect. Observational studies don't randomly assign individuals to groups, which rules out inference about cause and effect. An observational study that uses random sampling can make an inference about the population. The following example illustrates all four cases from the table above in a single setting.

## EXAMPLE

### *Vitamin C and Canker Sores*

#### Determining scope of inference

A small-town dentist wants to know if a daily dose of 500 milligrams (mg) of vitamin C will result in fewer canker sores in the mouth than taking no vitamin C.<sup>50</sup>

The dentist is considering the following four study designs.

*Design 1:* Get all dental patients in town with appointments in the next two weeks to take part in a study. Give each patient a survey with two questions: (1) Do you take at least 500 mg of vitamin C each day? (2) Do you frequently have canker sores? Based on patients' answers to Question 1, divide them into two groups: those who take at least 500 mg of vitamin C daily and those who don't.

*Design 2:* Get all dental patients in town with appointments in the next two weeks to take part in a study. Randomly assign half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months.

*Design 3:* Select a random sample of dental patients in town and get them to take part in a study. Divide the patients into two groups as in Design 1.

*Design 4:* Select a random sample of dental patients in town and get them to take part in a study. Randomly assign half of them to take 500 mg of vitamin C each day and the other half to abstain from taking vitamin C for three months.

For whichever design the dentist chooses, suppose she compares the proportion of patients in each group who complain of canker sores. Also suppose that she finds a statistically significant difference, with a smaller proportion of those taking vitamin C having canker sores.

**PROBLEM:** What can the dentist conclude for each design?

**SOLUTION:**

*Design 1:* Since the patients were not randomly selected, the dentist cannot infer that this result holds for a larger population of dental patients. This was an observational study since no treatments

were deliberately imposed on the patients. With no random assignment to the two groups, no inference about cause and effect can be made. The dentist just knows that for these patients, those who took vitamin C had fewer canker sores than those who didn't.

*Design 2:* As in Design 1, the dentist can't make any inference about this result holding for a larger population of dental patients. However, the treatments were randomly assigned to the subjects. Assuming proper control in the experiment, she can conclude that taking vitamin C reduced the chance of getting canker sores in her subjects.

*Design 3:* Since the patients were randomly selected from the population of dental patients in the town, the dentist can generalize the results of this study to the population. Because this was an observational study, no inference about cause and effect can be made. The dentist would conclude that for the population of dental patients in this town, those taking vitamin C have fewer canker sores than those who don't. She can't say whether the vitamin C causes this reduction or some other confounding variable. (Note that the dentist can't draw a conclusion about whether those people in the general population of the town who take vitamin C will have fewer canker sores, since the sample was only of dental patients.)

*Design 4:* As in Design 3, the random sampling allows the dentist to generalize the results of this study to the population of dental patients in the town. As in Design 2, the random assignment would allow her to conclude (assuming proper control in the experiment) that taking vitamin C reduced the chance of getting canker sores. So the dentist would conclude that for the population of dental patients in this town, those taking vitamin C will tend to have fewer canker sores than those who don't due to the vitamin C. As with Design 3, she can't extend this result to the general population of the town.

For Practice Try Exercise 105

## The Challenges of Establishing Causation

A well-designed experiment tells us that changes in the explanatory variable cause changes in the response variable. More precisely, it tells us that this happened for specific individuals in the specific environment of this specific experiment. The serious threat is that the treatments, the subjects, or the environment of our experiment may not be realistic. *Lack of realism* can limit our ability to apply the conclusions of an experiment to the settings of greatest interest.

### EXAMPLE

#### *Do Center Brake Lights Reduce Rear-End Crashes?*

##### Lack of realism

Do those high center brake lights, required on all cars sold in the United States since 1986, really reduce rear-end collisions? Randomized comparative experiments with fleets of rental and business cars, done before the lights were required, showed that the third brake light reduced rear-end collisions by as much as 50%. But requiring the third light in all cars led to only a 5% drop.

What happened? Most cars did not have the extra brake light when the experiments were carried out, so it caught the eye of following drivers. Now that almost all cars have the third light, they no longer capture attention.

In some cases, it isn't practical or even ethical to do an experiment. Consider these important questions:

- Does texting while driving increase the risk of having an accident?
- Does going to church regularly help people live longer?
- Does smoking cause lung cancer?

To answer these cause-and-effect questions, we just need to perform a randomized comparative experiment. Unfortunately, we can't randomly assign people to text while driving or to attend church or to smoke cigarettes. The best data we have about these and many other cause-and-effect questions come from observational studies.

It is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can be.

## EXAMPLE

### *Does Smoking Cause Lung Cancer?*

#### Living with observational studies

Doctors had long observed that most lung cancer patients were smokers. Comparison of smokers and similar nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be due to a lurking variable? Is there some genetic factor that makes people both more likely to get addicted to nicotine and to develop lung cancer? If so, then smoking and lung cancer would be strongly associated even if smoking had no direct effect on the lungs. Or maybe confounding is to blame. It might be that smokers live unhealthy lives in other ways (diet, alcohol, lack of exercise) and that some other habit confounded with smoking is a cause of lung cancer. How were these objections overcome?



What are the criteria for establishing causation when we can't do an experiment?

- *The association is strong.* The association between smoking and lung cancer is very strong.
- *The association is consistent.* Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.
- *Larger values of the explanatory variable are associated with stronger responses.* People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time.* Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.



- *The alleged cause is plausible.* Experiments with animals show that tars from cigarette smoke do cause cancer.

Medical authorities do not hesitate to say that smoking causes lung cancer. The U.S. Surgeon General states that cigarette smoking is “the largest avoidable cause of death and disability in the United States.”<sup>51</sup> The evidence for causation is overwhelming—but it is not as strong as the evidence provided by well-designed experiments. Conducting an experiment in which some subjects were forced to smoke and others were not allowed to would be unethical. In cases like this, observational studies are our best source of reliable information.

## Data Ethics\*

Medical professionals are taught to follow the basic principle “First, do no harm.” Shouldn’t those who carry out statistical studies follow the same principle? Most reasonable people think so. But this may not always be as simple as it sounds. Decide whether you think each of the following studies is ethical or unethical.

- A promising new drug has been developed for treating cancer in humans. Before giving the drug to human subjects, researchers want to administer the drug to animals to see if there are any potentially serious side effects.
- Are companies discriminating against some individuals in the hiring process? To find out, researchers prepare several equivalent résumés for fictitious job applicants, with the only difference being the gender of the applicant. They send the fake résumés to companies advertising positions and keep track of the number of males and females who are contacted for interviews.
- In a medical study of a new drug for migraine sufferers, volunteer subjects are randomly assigned to two groups. Members of the first group are given a placebo pill. Subjects in the second group are given the new drug. None of the subjects knows whether they are taking a placebo or the active drug. Neither do any of the physicians who are interacting with the subjects.
- Will people try to stop someone from driving drunk? A television news program hires an actor to play a drunk driver and uses a hidden camera to record the behavior of individuals who encounter the driver.

The most complex issues of data ethics arise when we collect data from people. The ethical difficulties are more severe for experiments that impose some treatment on people than for sample surveys that simply gather information. Trials of new medical treatments, for example, can do harm as well as good to their subjects. Here are some basic standards of data ethics that must be obeyed by all studies that gather data from human subjects, both observational studies and experiments.

### Basic Data Ethics

All planned studies must be reviewed in advance by an **institutional review board** charged with protecting the safety and well-being of the subjects.

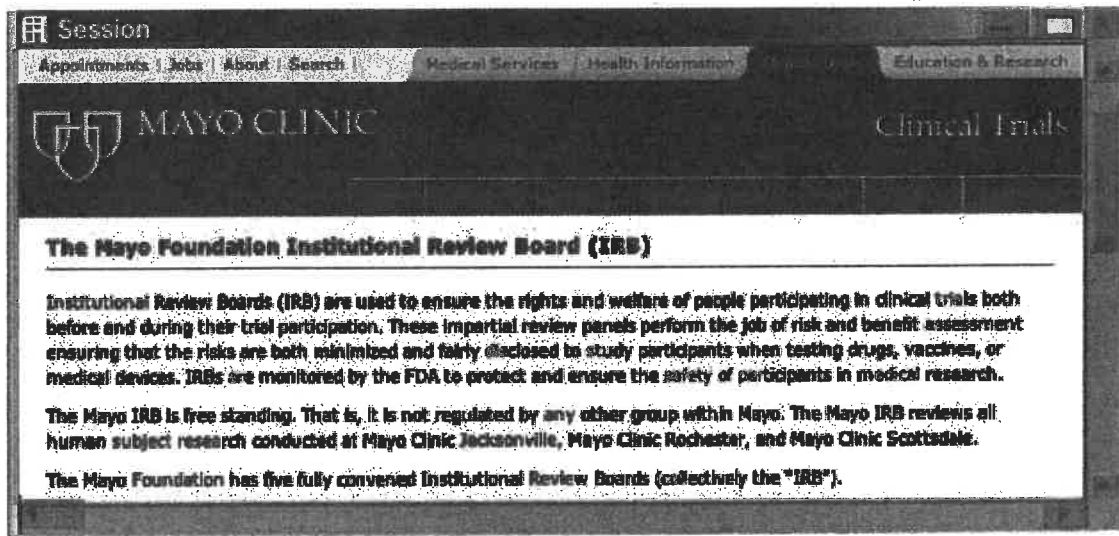
All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential**. Only statistical summaries for groups of subjects may be made public.

\*This is an important topic, but it is not required for the AP Statistics exam.

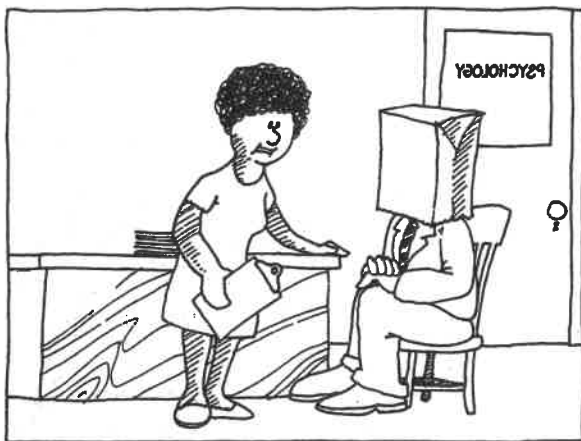
The law requires that studies carried out or funded by the federal government obey these principles.<sup>52</sup> But neither the law nor the consensus of experts is completely clear about the details of their application.

**Institutional review boards** The purpose of an institutional review board is not to decide whether a proposed study will produce valuable information or whether it is statistically sound. The board's purpose is, in the words of one university's board, "to protect the rights and welfare of human subjects (including patients) recruited to participate in research activities." The board reviews the plan of the study and can require changes. It reviews the consent form to be sure that subjects are informed about the nature of the study and about any potential risks. Once research begins, the board monitors its progress at least once a year.



The Web page of the Mayo Clinic's institutional review board. It begins by describing the job of such boards.

**Informed consent** Both words in the phrase "informed consent" are important, and both can be controversial. Subjects must be *informed* in advance about the nature of a study and any risk of harm it may bring. In the case of a sample survey, physical harm is not possible. But a survey on sensitive issues could result in emotional harm. The participants should be told what kinds of questions the survey will ask and about how much of their time it will take. Experimenters must tell subjects the nature and purpose of the study and outline possible risks. Subjects must then *consent* in writing.



"I realize the participants in this study are to be anonymous, but you're going to have to expose your eyes."

**Confidentiality** Ethical problems do not disappear once a study has been cleared by the review board, has obtained consent from its participants, and has actually collected data about them. It is important to protect individuals' privacy by keeping all data about them **confidential**. The report of an opinion poll may say what percent of the 1200 respondents felt that legal immigration should be reduced. It may not report what *you* said about this or any other issue.

Confidentiality is not the same as **anonymity**. Anonymity means that individuals are anonymous—their names

are not known even to the director of the study. Anonymity is rare in statistical studies. Even where anonymity is possible (mainly in surveys conducted by mail), it prevents any follow-up to improve nonresponse or inform individuals of results.

Any breach of confidentiality is a serious violation of data ethics. The best practice is to separate the identity of the study's participants from the rest of the data at once. A clever computer search of several data bases might be able, by combining information, to identify you and learn a great deal about you even if your name and other identification have been removed from the data available for search. Privacy and confidentiality of data are hot issues among statisticians in the computer age.

## • ACTIVITY *Response bias*

In this Activity, your team will design and conduct an experiment to investigate the effects of response bias in surveys.<sup>53</sup> You may choose the topic for your surveys, but you must design your experiment so that it can answer at least one of the following questions:

- Can the wording of a question create response bias?
- Do the characteristics of the interviewer create response bias?
- Does anonymity change the responses to sensitive questions?
- Does manipulating the answer choices change the response?

1. Write a proposal describing the design of your experiment. Be sure to include

- (a) your chosen topic and which of the above questions you'll try to answer.
- (b) a detailed description of how you will obtain your subjects (minimum of 50). Your plan must be practical!
- (c) what your questions will be and how they will be asked.
- (d) a clear explanation of how you will implement your design.
- (e) precautions you will take to collect data ethically.

Here are two examples of successful student projects:

*"Make-Up," by Caryn S. and Trisha T.* (all questions asked to males)

1. "Do you find females who wear makeup attractive?" (questioner wearing makeup: 75% answered yes)
2. "Do you find females who wear makeup attractive?" (questioner not wearing makeup: 30% answered yes)

*"Cartoons" by Sean W. and Brian H.*

1. "Do you watch cartoons?" (90% answered yes)
2. "Do you *still* watch cartoons?" (60% answered yes)

2. Once your teacher has approved your design, carry out the experiment. Record your data in tabular form.

3. Analyze your data. What conclusion do you draw? Provide appropriate graphical and numerical evidence to support your answer.

4. Prepare a report that includes the data you collected, your analysis from Step 3, and a discussion of any problems you encountered and how you dealt with them.

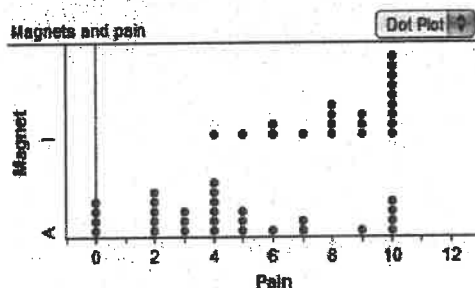
## case closed

*Magnets and Pain*

Let's return to the chapter-opening Case Study (page 205). Researchers carried out an experiment using as subjects 50 volunteer patients with chronic pain. A completely randomized design was used, with 29 subjects receiving the active-magnet treatment and 21 subjects getting the fake-magnet treatment. The Fathom dotplot shows the final pain ratings for both groups (A = active magnet, I = inactive magnet).

The average pain rating after the treatment was 8.43 for the inactive-magnet group and 4.38 for the active-magnet group. This difference is statistically significant. What conclusion can we draw?

The difference in the average responses for the two groups was too large to be explained by chance variation in the random assignment. So we can infer that the magnet treatment caused a reduction in pain. But since the subjects were volunteers, we can't safely generalize the findings of this experiment to any larger population.



## SECTION 4.3

## Summary

- Most statistical studies aim to make inferences that go beyond the data actually produced. **Inference about the population** requires that the individuals taking part in a study be randomly selected from the larger population. A well-designed experiment that randomly assigns treatments to experimental units allows **inference about cause and effect**.
- **Lack of realism** in an experiment can prevent us from generalizing its results.
- In the absence of an experiment, good evidence of causation requires a strong association that appears consistently in many studies, a clear explanation for the alleged causal link, and careful examination of possible lurking variables.
- Studies involving humans must be screened in advance by an **institutional review board**. All participants must give their **informed consent** before taking part. Any information about the individuals in the study must be kept **confidential**.
- Remember that randomized comparative experiments can answer questions that can't be answered without them. Also remember that "the interests of the subject must always prevail over the interests of science and society."<sup>54</sup>

## SECTION 4.3

## Exercises

102. **Random sampling versus random assignment**

Explain the difference between the types of inference that can be made as a result of random sampling and random assignment.

103. **Foster care versus orphanages** Do abandoned children placed in foster homes do better than similar children placed in an institution? The Bucharest Early Intervention Project found that the answer is a clear “Yes.” The subjects were 136 young children abandoned at birth and living in orphanages in Bucharest, Romania. Half of the children, chosen at random, were placed in foster homes. The other half remained in the orphanages.<sup>55</sup> (Foster care was not easily available in Romania at the time and so was paid for by the study.) What conclusion can we draw from this study? Explain.

104. **Frozen batteries** Will storing batteries in a freezer make them last longer? To find out, a company that produces batteries takes a random sample of 100 AA batteries from its warehouse. The company statistician randomly assigns 50 batteries to be stored in the freezer and the other 50 to be stored at room temperature for 3 years. At the end of that time period, each battery’s charge is tested. *Result:* Batteries stored in the freezer had a higher average charge, and the difference between the groups was statistically significant. What conclusion can we draw from this study? Explain.

105. **Who talks more—women or men?** According to Louann Brizendine, author of *The Female Brain*, women say nearly three times as many words per day as men. Skeptical researchers devised a study to test this claim. They used electronic devices to record the talking patterns of 396 university students who volunteered to participate in the study. The device was programmed to record 30 seconds of sound every 12.5 minutes without the carrier’s knowledge. According to a published report of the study in *Scientific American*, “Men showed a slightly wider variability in words uttered.... But in the end, the sexes came out just about even in the daily averages: women at 16,215 words and men at 15,669.”<sup>56</sup> This difference was not statistically significant. What conclusion can we draw from this study? Explain.

106. **Attend church, live longer?** One of the better studies of the effect of regular attendance at religious services gathered data from a random sample of

3617 adults. The researchers then measured lots of variables, not just the explanatory variable (religious activities) and the response variable (length of life). A news article said: “Churchgoers were more likely to be nonsmokers, physically active, and at their right weight. But even after health behaviors were taken into account, those not attending religious services regularly still were about 25% more likely to have died.”<sup>57</sup> What conclusion can we draw from this study? Explain.

107. **Daytime running lights** Canada requires that cars be equipped with “daytime running lights,” headlights that automatically come on at a low level when the car is started. Some manufacturers are now equipping cars sold in the United States with running lights. Will running lights reduce accidents by making cars more visible? An experiment conducted in a driving simulator suggests that the answer may be “Yes.” What concerns would you have about generalizing the results of such an experiment?

108. **Studying frustration** A psychologist wants to study the effects of failure and frustration on the relationships among members of a work team. She forms a team of students, brings them to the psychology lab, and has them play a game that requires teamwork. The game is rigged so that they lose regularly. The psychologist observes the students through a one-way window and notes the changes in their behavior during an evening of game playing. Can the psychologist generalize the results of her study to a team of employees that spends months developing a new product that never works right and is finally abandoned by their company? Explain.

- 109.\* **Minimal risk?** You have been invited to serve on a college’s institutional review board. You must decide whether several research proposals qualify for lighter review because they involve only minimal risk to subjects. Federal regulations say that “minimal risk” means the risks are no greater than “those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.” That’s vague. Which of these do you think qualifies as “minimal risk”?

\*Exercises 109 to 116: This is an important topic, but it is not required for the AP Statistics exam.



- (a) Draw a drop of blood by pricking a finger to measure blood sugar.
- (b) Draw blood from the arm for a full set of blood tests.
- (c) Insert a tube that remains in the arm, so that blood can be drawn regularly.
- 110.\* **Who reviews?** Government regulations require that institutional review boards consist of at least five people, including at least one scientist, one nonscientist, and one person from outside the institution. Most boards are larger, but many contain just one outsider.
- (a) Why should review boards contain people who are not scientists?
- (b) Do you think that one outside member is enough? How would you choose that member? (For example, would you prefer a medical doctor? A member of the clergy? An activist for patients' rights?)
- 111.\* **No consent needed?** In which of the circumstances below would you allow collecting personal information without the subjects' consent?
- (a) A government agency takes a random sample of income tax returns to obtain information on the average income of people in different occupations. Only the incomes and occupations are recorded from the returns, not the names.
- (b) A social psychologist attends public meetings of a religious group to study the behavior patterns of members.
- (c) A social psychologist pretends to be converted to membership in a religious group and attends private meetings to study the behavior patterns of members.
- 112.\* **Surveys of youth** A survey asked teenagers whether they had ever consumed an alcoholic beverage. Those who said "Yes" were then asked, "How old were you when you first consumed an alcoholic beverage?" Should consent of parents be required to ask minors about alcohol, drugs, and other such issues, or is consent of the minors themselves enough? Give reasons for your opinion.
- 113.\* **Anonymous? Confidential?** One of the most important nongovernment surveys in the United States is the National Opinion Research Center's General Social Survey. The GSS regularly monitors public opinion on a wide variety of political and social issues. Interviews are conducted in person in the subject's home. Are a subject's responses to GSS questions anonymous, confidential, or both? Explain your answer.
- 114.\* **Anonymous? Confidential?** Texas A&M, like many universities, offers screening for HIV, the virus that causes AIDS. Students may choose either anonymous
- or confidential screening. An announcement says, "Persons who sign up for screening will be assigned a number so that they do not have to give their name." They can learn the results of the test by telephone, still without giving their name. Does this describe the *anonymous* or the *confidential* screening? Why?
- 115.\* **The Willowbrook hepatitis studies** In the 1960s, children entering the Willowbrook State School, an institution for the mentally retarded, were deliberately infected with hepatitis. The researchers argued that almost all children in the institution quickly became infected anyway. The studies showed for the first time that two strains of hepatitis existed. This finding contributed to the development of effective vaccines. Despite these valuable results, the Willowbrook studies are now considered an example of unethical research. Explain why, according to current ethical standards, useful results are not enough to allow a study.
- 116.\* **Unequal benefits** Researchers on aging proposed to investigate the effect of supplemental health services on the quality of life of older people. Eligible patients on the rolls of a large medical clinic were to be randomly assigned to treatment and control groups. The treatment group would be offered hearing aids, dentures, transportation, and other services not available without charge to the control group. The review board felt that providing these services to some but not other persons in the same institution raised ethical questions. Do you agree?
117. **Animal testing (1.1)** "It is right to use animals for medical testing if it might save human lives." The General Social Survey asked 1152 adults to react to this statement. Here is the two-way table of their responses:

	Male	Female
Strongly agree	76	59
Agree	270	247
Neither agree nor disagree	87	139
Disagree	61	123
Strongly disagree	22	68

How do the distributions of opinion differ between men and women? Give appropriate graphical and numerical evidence to support your answer.

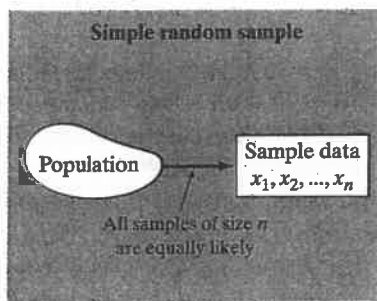
118. **Initial public offerings (1.3)** The business magazine *Forbes* reports that 4567 companies sold their first stock to the public between 1990 and 2000. The *mean* change in the stock price of these companies since the first stock was issued was +111%. The *median* change was -31%.<sup>58</sup> Explain how this could happen. (*Hint:* Start with the fact that Cisco Systems stock went up 60,600%.)

\*Exercises 109 to 116: This is an important topic, but it is not required for the AP Statistics exam.

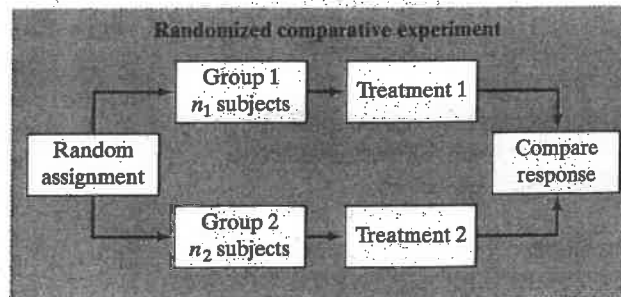
# Chapter 4 Chapter Summary

This chapter has explained good techniques for producing data and has also explained why bad techniques often produce worthless data. If we want to draw conclusions that go beyond the data at hand, then we have to design the data production appropriately. The figures below display the big ideas visually.

Random sampling and randomized comparative experiments are perhaps the two most important statistical inventions of the twentieth century. Unfortunately, you will still see many voluntary response samples and uncontrolled experiments.



(a)



(b)

## Chapter Review Exercises

*These exercises are designed to help you review the important ideas and methods of the chapter. Relevant learning objectives are provided in bulleted form before each exercise.*

- Identify the population and sample in a sample survey.
- Describe the scope of inference that is appropriate in an observational study.

**R4.1. Ontario Health Survey** The Ministry of Health in the province of Ontario, Canada, wants to know whether the national health care system is achieving its goals in the province. Much information about health care comes from patient records, but that source doesn't allow us to compare people who use health services with those who don't. So the Ministry of Health conducted the Ontario Health Survey, which interviewed a random sample of 61,239 people who live in the province of Ontario.<sup>59</sup>

- (a) What is the population for this sample survey? What is the sample?

- (b) The survey found that 76% of males and 86% of females in the sample had visited a general practitioner at least once in the past year. Do you think these estimates are close to the truth about the entire population? Why?

- Identify voluntary response samples and convenience samples. Explain how these bad sampling methods can lead to bias.

**R4.2. Bad sampling** A large high school wants to gather student opinion about parking for students on campus. It isn't practical to contact all students.

- (a) Give an example of a way to choose a voluntary response sample of students. Explain how this method could lead to bias.
- (b) Give an example of a way to choose a convenience sample of students. Explain how this method could lead to bias.

- Describe how to use Table D to select a simple random sample (SRS).

**R4.3. Drug testing** A baseball team regularly conducts random drug tests on its players. The 25 members of the team are listed below.

Agarwal	Chen	Healy	Moser	Roberts
Andrews	Frank	Hixson	Musselman	Shen
Baer	Fuest	Lee	Pavnica	Smith
Berger	Fuhrmann	Lynch	Petrucelli	Sundheim
Brockman	Garcia	Milhalko	Reda	Wilson

- Explain how you would use the line of random digits below to select an SRS of 3 team members for a random drug test.
- Use your method from (a) to choose the SRS. Show how you are using the digits below.

17521 78009 46239 84569 03316

- Distinguish a simple random sample from a stratified random sample or cluster sample. Give advantages and disadvantages of each sampling method.

**R4.4. Polling the faculty** A researcher wants to study the attitudes of college faculty members about the work habits of entering freshmen. These attitudes appear to differ depending on the type of college. The American Association of University Professors classifies colleges as follows:

Class I: Offer doctorate degrees and award at least 15 per year.

Class IIA: Award degrees above the bachelor's but are not in Class I.

Class IIB: Award no degrees beyond the bachelor's.

Class III: Two-year colleges.

The researcher would like to survey about 200 faculty members. Would you recommend a simple random sample, stratified random sample, or cluster sample? Justify your answer.

- Explain how undercoverage, nonresponse, and question wording can lead to bias in a sample survey.

**R4.5. Been to the movies?** An opinion poll calls 2000 randomly chosen residential telephone numbers, then asks to speak with an adult member of the household. The interviewer asks, "How many movies have you watched in a movie theater in the past 12 months?" In all, 1131 people respond.

- Identify a potential source of bias related to the question being asked. Suggest a change that would help fix this problem.
- Identify a potential source of bias in this survey that is not related to question wording. Suggest a change that would help fix this problem.

- Distinguish between an observational study and an experiment.
- Explain how a lurking variable in an observational study can lead to confounding.

**R4.6. Are anesthetics safe?** The National Halothane Study was a major investigation of the safety of anesthetics used in surgery. Records of over 850,000 operations performed in 34 major hospitals showed the following death rates for four common anesthetics:<sup>60</sup>

Anesthetic:	A	B	C	D
Death rate:	1.7%	1.7%	3.4%	1.9%

There seems to be a clear association between the anesthetic used and the death rate of patients. Anesthetic C appears to be dangerous.

- Explain why we call the National Halothane Study an observational study rather than an experiment, even though it compared the results of using different anesthetics in actual surgery.
- When the study looked at lurking variables that are related to a doctor's choice of anesthetic, it found that Anesthetic C was not causing extra deaths. Describe how a lurking variable could lead to confounding in this setting.

- Identify the experimental units or subjects, explanatory variables (factors), treatments, and response variables in an experiment.

**R4.7. Ugly fries** Few people want to eat discolored french fries. Potatoes are kept refrigerated before being cut for french fries to prevent spoiling and preserve flavor. But immediate processing of cold potatoes causes discoloring due to complex chemical reactions. The potatoes must therefore be brought to room temperature before processing. Researchers want to design an experiment in which tasters will rate the color and flavor of french fries prepared from several groups of potatoes. The potatoes will be freshly picked or stored for a month at room temperature or stored for a month refrigerated. They will then be sliced and cooked either immediately or after an hour at room temperature. Identify the experimental units, explanatory variable(s), the treatments, and the response variable(s).

- Describe a completely randomized design for an experiment.

**R4.8. Attitudes toward homeless people** Negative attitudes toward poor people are common. Are attitudes more negative when a person is homeless? To find out, read to subjects a description of a poor person. There are two versions. One begins