

Chapter 4



Collecting Data

Introduction	220
Section 4.1	221
Sampling and Surveys	
Section 4.2	241
Experiments	
Section 4.3	269
Using Studies Wisely	
Chapter 4 Wrap-Up	
Free Response AP [®] Problem, Yay!	285
Chapter 4 Review	285
Chapter 4 Review Exercises	287
Chapter 4 AP [®] Statistics Practice Test	289
Chapter 4 Project	292
Cumulative AP [®] Practice Test 1	292



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 88.5% in 2015 to 90.1% in 2016.¹
- According to a survey, U.S. teens aged 13 to 18 use entertainment media (television, Internet, social media, listening to music, etc.) nearly 9 hours a day, on average.²
- A study suggests that lack of sleep increases the risk of catching a cold.³
- For their final project, two AP[®] Statistics students showed that listening to music while studying decreased subjects' performance on a memory task.⁴

Can we trust these results? As you'll learn in this chapter, the answer 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 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' media habits. Common Sense Media conducted this survey using a random sample of 1399 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 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.

In Section 4.3, we revisit two key ideas from Sections 4.1 and 4.2: drawing conclusions about a population based on a random sample and drawing conclusions about cause and effect based on a randomized experiment. In both cases, we will focus on the role of randomization in our analysis.

SECTION 4.1**Sampling and Surveys****LEARNING TARGETS** *By the end of the section, you should be able to:*

- Identify the population and sample in a statistical study.
- Identify voluntary response sampling and convenience sampling and explain how these sampling methods can lead to bias.
- Describe how to select a simple random sample with technology or a table of random digits.
- Describe how to select a sample using stratified random sampling and cluster sampling, distinguish stratified random sampling from cluster sampling, and give an advantage of each method.
- Explain how undercoverage, nonresponse, question wording, and other aspects of a sample survey can lead to bias.

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 by conducting 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 pose the question to a **sample** chosen to represent the entire **population** of young drivers.

DEFINITION Population, Census, Sample

The **population** in a statistical study is the entire group of individuals we want information about. A **census** collects data from every individual in the population. A **sample** is a subset of individuals in the population from which we collect data.

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.

EXAMPLE**Sampling monitors and voters**
Populations and samples

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

- The quality control manager at a factory that produces computer monitors selects 10 monitors from the production line each hour. The manager inspects each monitor for defects in construction and performance.
- Prior to an election, a news organization surveys 1000 registered voters to predict which candidate will be elected as president.



Cancan Chu/Getty Images

SOLUTION:

- (a) The population is all the monitors produced in this factory that hour. The sample is the 10 monitors selected from the production line.
- (b) The population is all registered voters. The sample is the 1000 registered voters surveyed.

Because the sample came from 1 hour's production at this factory, the population is the monitors produced that hour in this factory—not all monitors produced in the world or even all monitors produced by this factory.

FOR PRACTICE, TRY EXERCISE 1

The Idea of a Sample Survey

We often draw conclusions about a population based on a sample. Have you ever tasted a sample of ice cream and ordered a cone because the sample tastes good? Because 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 decide *what population* we want to describe. The second step is to decide *what we want to measure*.

DEFINITION Sample survey

A **sample survey** is a study that collects data from a sample that is chosen to represent a specific population.

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 questions, like the news organization survey in the preceding example. We'll avoid this more restrictive terminology.

The final step in planning a sample survey is to decide how to choose a sample from the population. Here is an activity that illustrates the process of conducting a sample survey.

ACTIVITY

Who wrote the Federalist Papers?

The Federalist Papers are a series of 85 essays supporting the ratification of the U.S. Constitution. At the time they were published, the identity of the authors was a secret known to only a few people. Over time, however, the authors were identified as Alexander Hamilton, James Madison, and John Jay. The authorship of 73 of the essays is fairly certain, leaving 12 in dispute. However, thanks in some part to statistical analysis, most scholars now believe that the 12 disputed essays were written by Madison alone or in collaboration with Hamilton.⁵

There are several ways to use statistics to help determine the authorship of a disputed text. One method is to estimate the average word length in a disputed text and compare it to the average word lengths of works where the authorship is not in dispute.



The following passage is the opening paragraph of Federalist Paper No. 51,⁶ one of the disputed essays. The theme of this essay is the separation of powers between the three branches of government.

To what expedient, then, shall we finally resort, for maintaining in practice the necessary partition of power among the several departments, as laid down in the Constitution? The only answer that can be given is, that as all these exterior provisions are found to be inadequate, the defect must be supplied, by so contriving the interior structure of the government as that its several constituent parts may, by their mutual relations, be the means of keeping each other in their proper places. Without presuming to undertake a full development of this important idea, I will hazard a few general observations, which may perhaps place it in a clearer light, and enable us to form a more correct judgment of the principles and structure of the government planned by the convention.

1. Choose 5 words from this passage. Count the number of letters in each of the words you selected, and find the average word length.
2. Your teacher will draw and label a horizontal axis for a class dotplot. Plot the average word length you obtained in Step 1 on the graph.
3. Your teacher will show you how to use a random number generator to select a random sample of 5 words from the 130 words in the opening passage. Count the number of letters in each of the selected words, and find the average word length.
4. Your teacher will draw and label another horizontal axis with the same scale for a comparative dotplot. Plot the average word length you obtained in Step 3 on the graph.
5. How do the dotplots compare? Can you think of any reasons why they might be different? Discuss with your classmates.

How to Sample Badly

Suppose we want to know how long students at a large high school spent doing homework last week. We might go to the school library and ask the first 30 students we see about the amount of time they spend on their homework. This method is known as a **convenience sampling**.

DEFINITION Convenience sampling

Convenience sampling selects individuals from the population who are easy to reach.



Convenience sampling often produces unrepresentative data. Consider our sample of 30 students from the school library. It's unlikely that this convenience sample accurately represents the homework habits of all students at the high

school. In fact, if we were to repeat this sampling process day after day, we would almost always overestimate the average homework time in the population. Why? Because students who hang out in the library tend to be more studious. This is **bias**: using a method that favors some outcomes over others.

DEFINITION Bias

The design of a statistical study shows **bias** if it is very likely to underestimate or very likely to overestimate the value you want to know.

AP® EXAM TIP

If you're asked to describe how the design of a sample survey leads to bias, you're expected to do two things: (1) describe how the members of the sample might respond differently from the rest of the population, and (2) explain how this difference would lead to an underestimate or overestimate. Suppose you were asked to explain how using your statistics class as a sample to estimate the proportion of all high school students who own a graphing calculator could result in bias. You might respond, "This is a convenience sample. It 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. So this method would probably lead to an overestimate of the actual population proportion."



Bias is not just bad luck in one sample. It's the result of a bad study design that will consistently miss the truth about the population in the same way. Convenience sampling will almost always result in bias. So will **voluntary response sampling**.

The Internet brings voluntary response sampling to the computer nearest you. Visit 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 much about the views of any larger population.

DEFINITION Voluntary response sampling

Voluntary response sampling allows people to choose to be in the sample by responding to a general invitation.

Most Internet polls, along with call-in, text-in, and write-in polls, rely on voluntary response sampling. *People who self-select to participate in such surveys are usually not representative of some larger population of interest.* Voluntary response sampling attracts people who feel strongly about an issue, and who often share the same opinion. That leads to bias.

EXAMPLE

Boaty McBoatface Biased sampling methods

PROBLEM: In 2016, Britain's Natural Environment Research Council invited the public to name its new \$300 million polar research ship. To vote on the name, people simply needed to visit a website and record their choice. Ignoring names suggested by the council, over 124,000 people voted for "Boaty McBoatface," which ended up having more than 3 times as many votes as the second-place finisher.⁷



NERC/SplashNews/Newscom

What type of sampling did the council use in their poll? Explain how bias in this sampling method could have affected the poll results.

SOLUTION:

The council used **voluntary response sampling**: people chose to go online and respond. The people who chose to be in the sample were probably less serious about science than the British population as a whole—and more likely to prefer a funny name. The proportion of people in the sample who prefer the name *Boaty McBoatface* is likely to be greater than the proportion of all British residents who would choose this name.

Remember to describe how the responses from the members of the sample might differ from the responses from the rest of the population *and* how this difference will affect the estimate.

FOR PRACTICE, TRY EXERCISE 5

**CHECK YOUR UNDERSTANDING**

For each of the following situations, identify the sampling method used. Then explain how bias in the sampling method could affect the results.

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 if 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 to respond “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 convenience sampling, the researcher chooses easy-to-reach members of the population. In voluntary response sampling, people decide whether to join the sample. Both sampling methods suffer from bias due to personal choice. As you discovered in The Federalist Papers activity, a good way to avoid bias is to let chance choose the sample. That’s the idea of **random sampling**.

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

DEFINITION Random sampling

Random sampling involves using a chance process to determine which members of a population are included in the sample.

For example, to choose a random sample of 6 students from a class of 30, start by writing each of the 30 names on a separate slip of paper, making sure the slips are all the same size. Then put the slips in a hat, mix them well, and pull out slips one at a time until you have identified 6 different students. An alternative approach would be to give each member of the population a distinct number and to use the “hat method” with these numbers instead of people’s names. Note that this version would work just as well if the population consisted of animals or things. The resulting sample is called a **simple random sample**, or **SRS** for short.

DEFINITION Simple random sample (SRS)

A **simple random sample (SRS)** of size n is chosen in such a way that every group of n individuals in the population has an equal chance to be selected as the sample.

An SRS gives every possible sample of the desired size an equal chance to be chosen. Picture drawing 20 slips of paper (the sample) from a hat containing 200 identical slips (the population). Any set of 20 slips has the same chance as any other set of 20 to be chosen. This also means that each individual has the same chance to be chosen in an SRS. However, giving each individual the same chance to be selected is not enough to guarantee that a sample is an SRS. Some other random sampling methods give each member of the population an equal chance to be selected, but not each possible sample. We'll look at some of these methods later.

HOW TO CHOOSE A SIMPLE RANDOM SAMPLE The hat method won't work well if the population is large. Imagine trying to take a simple random sample of 1000 registered voters in the United States using a hat! In practice, most people use random numbers generated by technology to choose samples.

HOW TO CHOOSE AN SRS WITH TECHNOLOGY

- **Label.** Give each individual in the population a distinct numerical label from 1 to N , where N is the number of individuals in the population.
- **Randomize.** Use a random number generator to obtain n *different* integers from 1 to N , where n is the sample size.
- **Select.** Choose the individuals that correspond to the randomly selected integers.

When choosing an SRS, we make the selections *without replacement*. That is, once an individual is selected for a sample, that individual cannot be selected again. Many random number generators sample numbers *with* replacement, so it is important to explain that repeated numbers should be ignored when using technology to select an SRS.

11. Technology Corner**CHOOSING AN SRS**

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

Let's use a graphing calculator to select an SRS of 10 students from a population of students numbered 1 to 1750.

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

- Press **MATH**, then select PROB (PRB) and choose randInt(.

Newer OS: In the dialogue box, enter these values: lower: 1, upper: 1750, n: 1, choose Paste, and press **ENTER**.

Older OS: Complete the command randInt(1,1750) and press **ENTER**.

- Compare your results with those of your classmates. If several students got the same number, you'll need to seed your calculator's random integer generator with different numbers before you proceed. Directions for doing this are given in the *Teacher's Edition*.
2. Randomly generate 10 distinct numbers from 1 to 1750 by pressing **ENTER** until you have chosen 10 different labels.

Note: If you have a TI-84 Plus CE, use the command `RandIntNoRep(1,1750,10)` to get 10 distinct integers from 1 to 1750. If you have a TI-84 with OS 2.55 or later, use the command `RandIntNoRep(1,1750)` to sort the numbers from 1 to 1750 in random order. The first 10 numbers listed give the labels of the chosen students.

NORMAL FLOAT AUTO REAL RADIAN CL	
<code>randInt(1,1750)</code>	139
<code>randInt(1,1750)</code>	1126
<code>randInt(1,1750)</code>	920
<code>randInt(1,1750)</code>	1089

There are many random number generators available on the Internet, including those at www.random.org. You can also use the random number generator on your calculator.

If you don't have technology handy, you can use a table of random digits to choose an SRS. We have provided a table of random digits at the back of the book (Table D). Here is an excerpt:

Table D Random digits

LINE								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056

You can think of this table as the result of someone putting the digits 0 to 9 in a hat, mixing, drawing one, replacing it, mixing again, drawing another, and so on. The digits have been arranged in groups of five within numbered rows to make the table easier to read. The groups and rows have no special meaning—Table D is just a long list of

randomly chosen digits. As with technology, there are three steps in using Table D to choose a random sample.

HOW TO CHOOSE AN SRS USING TABLE D

- **Label.** Give each member of the population a distinct numerical label with the same number of digits. Use as few digits as possible.
- **Randomize.** Read consecutive groups of digits of the appropriate length from left to right across a line in Table D. Ignore any group of digits that wasn't used as a label or that duplicates a label already in the sample. Stop when you have chosen n different labels.
- **Select.** Choose the individuals that correspond to the randomly selected integers.

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 or 0001, 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. Here's an example that shows how this process works.

EXAMPLE**Attendance audit****Choosing an SRS with Table D**

PROBLEM: Each year, the state Department of Education randomly selects three schools from each district and conducts a detailed audit of their attendance records.

- (a) Describe how to use a table of random digits to select an SRS of three schools from this list of 19 schools.

Amphitheater High School	Keeling Elementary School
Amphitheater Middle School	La Cima Middle School
Canyon del Oro High School	Mesa Verde Elementary School
Copper Creek Elementary School	Nash Elementary School
Coronado K-8 School	Painted Sky Elementary School
Cross Middle School	Prince Elementary School
Donaldson Elementary School	Rio Vista Elementary School
Harelson Elementary School	Walker Elementary School
Holaway Elementary School	Wilson K-8 School
Ironwood Ridge High School	



Holly Albrecht

- (b) Use the random digits here to choose the sample.

62081 64816 87374 09517 84534 06489 87201 97245

SOLUTION:

- (a) Label the schools from 01 to 19 in alphabetical order. Move along a line of random digits from left to right, reading two-digit numbers, until three different numbers from 01 to 19 have been selected (ignoring repeated numbers and the numbers 20–99, 00). Audit the three schools that correspond with the numbers selected.

Remember to include all three steps:

- Label
- Randomize
- Select

- (b) 62-skip, 08-select, 16-select, 48-skip, 16-repeat, 87-skip, 37-skip, 40-skip, 95-skip, 17-select. The three schools are 08: Harelson Elementary School, 16: Prince Elementary School, and 17: Rio Vista Elementary School.

FOR PRACTICE, TRY EXERCISE 11

**CHECK YOUR UNDERSTANDING**

A furniture maker buys hardwood in batches that each contain 1000 pieces. 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. Describe how to select a simple random sample of 5 pieces using each of the following.

1. A random number generator
2. A table of random digits

One of the most common alternatives to simple random sampling is called **stratified random sampling**. This method involves sampling from groups (**strata**) of similar individuals within the population separately. Then these separate “subsamples” are combined to form the sample.

Strata are groups of individuals in a population who share characteristics thought to be associated with the variables being measured in a study. **Stratified random sampling** selects a sample by choosing an SRS from each stratum and combining the SRSs into one overall sample.

Stratified random sampling works best when the individuals within each stratum are similar with respect to what is being measured and when there are large differences between strata. 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. After all, it is reasonable to think that freshmen have different sleep habits than seniors. The following activity illustrates the benefit of choosing appropriate strata.

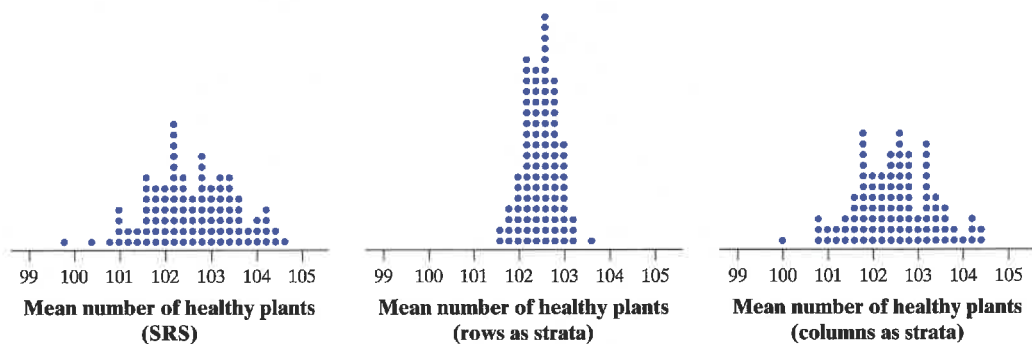
Sampling sunflowers



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. 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 (e.g., B6) of each square you select.
2. This time, select 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—for example, Row 1: G, Row 2: B, and so on.
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—for example, Column A: 4, Column B: 1, and so on.
4. Your teacher will provide the actual number of healthy sunflowers in each grid square. Use that information to calculate your estimate of the mean number of healthy sunflowers per square in the entire field for each of your samples in Steps 1, 2, and 3.
5. Make comparative dotplots showing the mean number of healthy sunflowers obtained using the three different sampling methods for all members of the class. Describe any similarities and differences you see.

The following dotplots show the mean number of healthy plants in 100 samples using each of the three sampling methods in the activity: simple random sampling, stratified random sampling with rows of the field as strata, and stratified random sampling with columns of the field as strata. Notice that all three distributions are centered at about 102.5, the true mean number of healthy plants in all squares of the field. That makes sense because random sampling tends to yield accurate estimates of unknown population means.



One other detail stands out in the graphs: there is much less variability in the estimates when we use the rows as strata. The table provided by your teacher shows the actual number of healthy sunflowers in each grid square. Notice that the squares within each row contain a similar number of healthy plants but that there are big differences between rows. *When we can choose strata that have similar responses (e.g., number of healthy plants) within strata but different responses between strata, stratified random samples give more precise estimates than simple random samples of the same size.*

Why didn't using the columns as strata reduce the variability of the estimates in a similar way? Because the numbers of healthy plants vary a lot within each column and aren't very different from other columns.

Both simple random sampling and stratified random sampling 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 (**clusters**) of individuals that are "near" one another. That's the idea of **cluster sampling**.

DEFINITION Clusters, Cluster sampling

A **cluster** is a group of individuals in the population that are located near each other. **Cluster sampling** selects a sample by randomly choosing clusters and including each member of the selected clusters in the sample.

Cluster sampling is often used for practical reasons, like saving time and money. It works best when the clusters look just like the population but on a smaller scale. Imagine a large high school that assigns students to homerooms alphabetically by last name, in groups of 25. Administrators want to survey 200 randomly selected students about a proposed schedule change. It would be difficult to track down an SRS of 200 students, so the administration opts for a cluster sample of homerooms. The principal (who knows some statistics) selects an SRS of 8 homerooms and gives the survey to all 25 students in each homeroom.

Be sure you understand the difference between strata and clusters. We want each stratum to contain similar individuals and for large differences to exist between strata. For a cluster sample, we'd like each cluster to look just like the population, but on a smaller scale. Unfortunately, cluster samples don't offer the statistical advantage of better information about the population that stratified random samples do. Here's an example that compares stratified random sampling and cluster sampling.

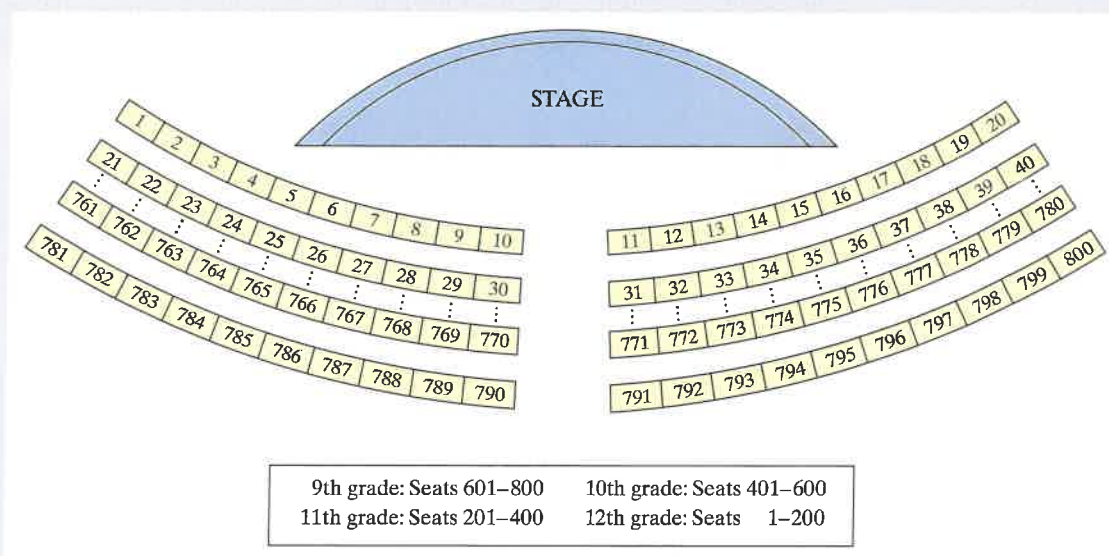
EXAMPLE

Sampling at a school assembly Other sampling methods

PROBLEM: The student council wants to conduct a survey about use of the school library during the first five minutes of an all-school assembly in the auditorium. There are 800 students present at the assembly. Here is a map of the auditorium. Note that students are seated by grade level and that the seats are numbered from 1 to 800.



H. Mark Weidman Photography/Alamy



- Describe how to obtain a sample of 80 students using stratified random sampling. Explain your choice of strata and why this method might be preferred to simple random sampling.
- Describe how to obtain a sample of 80 students using cluster sampling. Explain your choice of clusters and why this method might be preferred to simple random sampling.

SOLUTION:

- (a) Because students' library use might be similar within grade levels but different across grade levels, use the grade-level seating areas as strata. For the 9th grade, generate 20 different random integers from 601 to 800 and give the survey to the students in those seats. Do the same for sophomores, juniors, and seniors using their corresponding seat numbers. Stratification by grade level should result in more precise estimates of student opinion than a simple random sample of the same size.
- (b) Each column of seats from the stage to the back of the auditorium could be used as a cluster because it would be relatively easy to hand out the surveys to an entire column. Label the columns from 1 to 20 starting at the left side of the stage, generate 2 different integers from 1 to 20, and give the survey to the 80 students sitting in these two columns. Cluster sampling is much more efficient than finding 80 seats scattered about the auditorium, as required by simple random sampling.

Note that each cluster contains students from all four grade levels, so each should represent the population fairly well. Randomly selecting 4 rows as clusters would also be easy, but this may over- or under-represent one grade level.

FOR PRACTICE, TRY EXERCISE 21

Most large-scale sample surveys use *multistage sampling*, which combines two or more sampling methods. For example, the U.S. Census Bureau carries out a monthly Current Population Survey (CPS) of about 60,000 households. Researchers start by choosing a stratified random sample of neighborhoods in 756 of the 2007 geographical areas in the United States. Then they divide each neighborhood into clusters of four nearby households and select a cluster sample to interview.

Analyzing data from sampling methods other than simple random sampling 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

A factory runs 24 hours a day, producing wood pencils on three 8-hour shifts — day, evening, and overnight. In the last stage of manufacturing, the pencils are packaged in boxes of 10 pencils each. Each day a sample of 300 pencils is selected and inspected for quality.

1. Describe how to select a stratified random sample of 300 pencils. Explain your choice of strata.
2. Describe how to select a cluster sample of 300 pencils. Explain your choice of clusters.
3. Explain a benefit of using a stratified random sample and a benefit of using a cluster sample in this context.

Sample Surveys: What Else Can Go Wrong?

As we have learned, the use of bad sampling methods (convenience or voluntary response) often leads to bias. Researchers can avoid these methods by using random sampling to choose their samples. Other problems in conducting sample surveys are more difficult to avoid.

Sampling is sometimes done using a list of individuals in the population, called a sampling frame. Such lists are seldom accurate or complete. The result is **undercoverage**.

DEFINITION Undercoverage

Undercoverage occurs when some members of the population are less likely to be chosen or cannot be chosen in a sample.

Most samples suffer from some degree of undercoverage. A sample survey of households, for example, will miss not only homeless people but also prison inmates and students in dormitories. An opinion poll conducted by calling land-line telephone numbers will miss households that have only cell phones as well as households without a phone. The results of sample surveys may not be accurate if the people who are undercovered differ from the rest of the population in ways that affect their responses.

Even if every member of the population is equally likely to be selected for a sample, not all members of the population are equally likely to provide a response. Some people are never at home and cannot be reached by pollsters on the phone or in person. Other people see an unfamiliar phone number on their caller ID and never pick up the phone or quickly hang up when they don't recognize the voice of the caller. These are examples of **nonresponse**, another major source of bias in surveys.

DEFINITION Nonresponse

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

Nonresponse leads to bias when the individuals who can't be contacted or refuse to participate would respond differently from those who do participate. Consider a telephone survey that asks people how many hours of television they watch per day. People who are selected but are out of the house won't be able to respond. Because these people probably watch less television than the people who are at home when the phone call is made, the mean number of hours obtained in the sample is likely to be greater than the mean number of hours of TV watched in the population.

How bad is nonresponse? According to polling guru Nate Silver, "Response rates to political polls are dismal. Even polls that make every effort to contact a representative sample of voters now get no more than 10 percent to complete their surveys — down from about 35 percent in the 1990s."⁸ In contrast, 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%.⁹



Some students misuse the term *voluntary response* to explain why certain individuals don't respond in a sample survey. Their belief is that participation in the survey is optional (voluntary), so anyone can refuse to take part. What the students are 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.

The wording of questions has an important influence on the answers given to a sample survey. Confusing or leading questions can introduce strong bias. Even a single word can make a difference. In a recent Quinnipiac University poll, half of the respondents were asked if they support “stronger gun laws” and the other half were asked if they support “stronger gun control laws.” In the first group, 52% of respondents supported stronger laws, but when the word *control* was added to the question, only 46% of respondents supported stronger laws.¹⁰

The gender, age, ethnicity, or behavior of the interviewer can also affect people’s responses. People may lie about their age, income, or drug use. They may misremember how many hours they spent on the Internet last week. Or they might make up an answer to a question that they don’t understand. All these issues can lead to **response bias**.

DEFINITION Response bias

Response bias occurs when there is a systematic pattern of inaccurate answers to a survey question.

EXAMPLE

Wash your hands!

Response bias

PROBLEM: What percent of Americans wash their hands after using the bathroom? It depends on how you collect the data. In a telephone survey of 1006 U.S. adults, 96% said they always wash their hands after using a public restroom. An observational study of 6028 adults in public restrooms told a different story: only 85% of those observed washed their hands after using the restroom. Explain why the results of the two studies are so different.¹¹

SOLUTION:

When asked in person, many people may lie about always washing their hands because they want to appear to have good hygiene. When people are only observed and not asked directly, the percent who wash their hands will be smaller—and much closer to the truth.

FOR PRACTICE, TRY EXERCISE 29

Even the order in which questions are asked is important. For example, ask a sample of college students these two questions:

- “How happy are you with your life in general?” (Answer 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. The lesson is clear: the order in which questions are asked can influence the results.



CHECK YOUR UNDERSTANDING

- Each of the following is a possible source of bias in a sample survey. Name the type of bias that could result.
 - The sample is chosen at random from a telephone directory.
 - Some people cannot be contacted in five calls.
 - Interviewers choose people walking by on the sidewalk to interview.
- 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?¹²

Do you think the estimate of 84% is less than, greater than, or about equal to the percent of all people in the population who would oppose banning disposable diapers? Explain your reasoning.

Section 4.1 Summary

- A **census** collects data from every individual in the **population**.
- A **sample survey** selects a **sample** from the population of all individuals about which we desire information. The goal of a sample survey is to draw conclusions about the population based on data from the sample.
- Convenience sampling** chooses individuals who are easiest to reach. In **voluntary response sampling**, individuals choose to join the sample in response to an open invitation. Both these sampling methods usually lead to **bias**: they will be very likely to underestimate or very likely to overestimate the value you want to know.
- Random sampling** uses a chance process to select a sample.
- A **simple random sample (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 a table of random digits or technology to select the sample.
- To use **stratified random sampling**, divide the population into groups of individuals (**strata**) that are similar in some way that might affect their responses. Then choose a separate SRS from each stratum and combine these SRSs to form the sample. When strata are “similar within but different between,” stratified random samples tend to give more precise estimates of unknown population values than do simple random samples.
- To use **cluster sampling**, divide the population into groups of individuals that are located near each other, called **clusters**. Randomly select some of these clusters. All the individuals in the chosen clusters are included in the sample. Ideally, clusters are “different within but similar between.” Cluster sampling saves time and money by collecting data from entire groups of individuals that are close together.

- Random sampling helps avoid bias in choosing a sample. Bias can still occur in the sampling process due to **undercoverage**, which happens when some members of the population are less likely to be chosen or cannot be chosen for the sample.
- Other serious problems in sample surveys can occur after the sample is chosen. The single biggest problem is **nonresponse**: when people can't be contacted or refuse to answer. Untruthful answers by respondents, poorly worded questions, and other problems can lead to **response bias**.

4.1 Technology Corners

TI-Nspire and other technology instructions are on the book's website at highschool.bfwpub.com/tps6e.

11. Choosing an SRS

Page 226

Section 4.1

Exercises

- 1. 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 at random from the 1000 stuffed each hour for visual inspection. 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 a number to it. 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. Students as customers** A high school's student newspaper plans to survey local businesses about the importance of students as customers. From an alphabetical list of all local businesses, the newspaper staff chooses 150 businesses at random. Of these, 73 return the questionnaire mailed by the staff. 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.
- 5. Sleepless nights** How much sleep do high school students get on a typical school night? A counselor designed a survey to find out. To make data collection easier, the counselor 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 counselor obtain?
 - (b) Explain why this sampling method is biased. Is 7.2 hours probably greater than or less than the true average amount of sleep last night for all students at the school? Why?
- 6. Online polls** *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 subsequent 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 greater than or less than the true percent of all adults who believe that all cell phone use while driving should be banned? Why?
- 7. Online reviews** Many websites include customer reviews of products, restaurants, hotels, and so on. The manager of a hotel was upset to see that 26% of reviewers on a travel website gave the hotel "1 star"—the lowest possible rating. Explain how bias in the sampling method could affect the estimate.

8. Funding for fine arts The band director at a high school wants to estimate the percentage of parents who support a decrease in the budget for fine arts. Because many parents attend the school's annual musical, the director surveys the first 30 parents who arrive at the show. Explain how bias in the sampling method could affect the estimate.

9. 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?

10. Sampling mall shoppers You may have seen the mall interviewer, clipboard in hand, approaching people passing by. Explain why even a large sample of mall shoppers would not provide a trustworthy estimate of the current unemployment rate in the city where the mall is located.

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.

- (a) Explain how you would use a line of Table D to choose an SRS of 5 students from the following list.
- (b) Use line 107 to select the sample. Show how you use each of the digits.

Anderson	Drasin	Kim	Rider
Arroyo	Eckstein	Molina	Rodriguez
Batista	Fernandez	Morgan	Samuels
Bell	Fullmer	Murphy	Shen
Burke	Gandhi	Nguyen	Tse
Cabrera	Garcia	Palmiero	Velasco
Calloway	Glaus	Percival	Wallace
Delluci	Helling	Prince	Washburn
Deng	Husain	Puri	Zabidi
De Ramos	Johnson	Richards	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 following list.
- (b) Use line 117 to select the sample. Show how you use each of the digits.

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

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 the plots.¹³

- (a) Explain how you would use a random number generator to choose an SRS of 141 plots. Your description should be clear enough for a classmate to carry out your plan.
- (b) Use your method from part (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.¹⁴

- (a) Explain how you would use a random number generator to choose the SRS. Your description should be clear enough for a classmate to carry out your plan.
- (b) Use your method from part (a) to choose the first 3 gravestones.

15. 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 this area that have been infested.

- (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 exactly 35% of *all* the pine trees on the west side of the park are infested? Why or why not?
16. **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. Because 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 a different sampling method: Randomly choose one of the first 50 iPhones produced. Inspect that phone and every fiftieth iPhone produced afterward. (This method is known as *systematic random sampling*.) Explain carefully why this sampling method is *not* an SRS.
17. **No tipping** The owner of a large restaurant is considering a new "no tipping" policy and wants to survey a sample of employees. The policy would add 20% to the cost of food and beverages and the additional revenue would be distributed equally among servers and kitchen staff. Describe how to select a stratified random sample of approximately 30 employees. Explain your choice of strata and why stratified random sampling might be preferred in this context.
18. **Parking on campus** The director of student life at a university wants to estimate the proportion of undergraduate students who regularly park a car on campus. Describe how to select a stratified random sample of approximately 100 students. Explain your choice of strata and why stratified random sampling might be preferred in this context.
19. **SRS of engineers?** A corporation employs 2000 male and 500 female engineers. A stratified random sample of 200 male and 50 female engineers gives every individual in the population the same chance to be chosen for the sample. Is it an SRS? Explain your answer.
20. **SRS of students?** 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 their attitudes toward alcohol. You have given every student at the party the same chance to be interviewed. Is your sample an SRS? Explain your answer.


21. How is your room? 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 select 120 rooms and survey the registered guest in each of the selected rooms about his or her overall satisfaction with the property.

- (a) Describe how to obtain a sample of 120 rooms using stratified random sampling. Explain your choice of strata and why this method might be preferred to simple random sampling.
- (b) Describe how to obtain a sample of 120 rooms using cluster sampling. Explain your choice of clusters and why this method might be preferred to simple random sampling.

22. **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 wants to survey fans about concessions that are sold during games. Tickets are most expensive for seats on the sidelines. The cheapest seats are in the end zones (where one of the authors sat as a student). A map of the stadium is shown.



- (a) Describe how to obtain a sample using stratified random sampling. Explain your choice of strata and why this method might be preferred to simple random sampling.
- (b) Describe how to obtain a sample using cluster sampling. Explain your choice of clusters and why this method might be preferred to simple random sampling.

- 23. High-speed Internet** Laying fiber-optic cable is expensive. Cable companies want to make sure that if they extend their lines to less dense suburban or rural areas, there will be sufficient demand so 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.
- What is the name for this kind of sampling method?
 - Give a possible reason why the cable company chose this method.
- 24. Timber!** A lumber company wants to estimate the proportion of trees in a large forest that are ready to be cut down. They use an aerial map to divide the forest into 200 equal-sized rectangles. Then they choose a random sample of 20 rectangles and examine every tree that's in one of those rectangles.
- What is the name for this kind of sampling method?
 - Give a possible reason why the lumber company chose this method.
- 25. Eating on campus** The director of student life at a small college wants to know what percent of students eat regularly in the cafeteria. To find out, the director selects an SRS of 300 students who live in the dorms. Describe how undercoverage might lead to bias in this study. Explain the likely direction of the bias.
- 26. Immigration reform** A news organization wants to know what percent of U.S. residents support a "pathway to citizenship" for people who live in the United States illegally. The news organization randomly selects registered voters for the survey. Describe how undercoverage might lead to bias in this study. Explain the likely direction of the bias.
- 27. Reporting weight loss** A total of 300 people participated in a free 12-week weight-loss course at a community health clinic. After one year, administrators emailed each of the 300 participants to see how much weight they had lost since the end of the course. Only 56 participants responded to the survey. The mean weight loss for this sample was 13.6 pounds. Describe how nonresponse might lead to bias in this study. Explain the likely direction of the bias.
- 28. Nonresponse** A survey of drivers began by randomly sampling from 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.¹⁵ Describe how nonresponse might lead to bias in this study. Explain the likely direction of the bias.
- 29. Running red lights** An SRS of 880 drivers was asked:  **pg 234** "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. Explain the likely direction of the bias.
- 30. 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.¹⁶ Explain why the two percentages are so different.
- 31. Boys don't cry?** Two female statistics students asked a random sample of 60 high school boys if they have ever cried during a movie. Thirty of the boys were asked directly and the other 30 were asked anonymously by means of a "secret ballot." When the responses were anonymous, 63% of the boys said "Yes," whereas only 23% of the other group said "Yes." Explain why the two percentages are so different.
- 32. Weight? Wait what?** Marcos asked a random sample of 50 mall shoppers for their weight. Twenty-five of the shoppers were asked directly and the other 25 were asked anonymously by means of a "secret ballot." The mean reported weight was 13 pounds heavier for the anonymous group. Explain why the two means are so different.¹⁷
- 33. 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?
- "Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?"
 - "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?"
 - "In view of escalating environmental degradation and incipient resource depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?"
- 34. 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?
- Which of the following best represents your opinion on gun control?
 - The government should confiscate our guns.
 - 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 35–40.

35. A popular website places opinion poll questions next to many of its news stories. Simply click your response to join the sample. One of the questions was “Do you plan to diet this year?” More than 30,000 people responded, with 68% saying “Yes.” Which of the following is true?

- (a) About 68% of Americans planned to diet.
- (b) The poll used a convenience sample, so the results tell us little about the population of all adults.
- (c) The poll uses voluntary response, so the results tell us little about the population of all adults.
- (d) The sample is too small to draw any conclusion.
- (e) None of these.

36. To gather information about the validity of a new standardized test for 10th-grade students in a particular state, a random sample of 15 high schools was selected from the state. The new test was administered to every 10th-grade student in the selected high schools. What kind of sample is this?

- (a) A simple random sample
- (b) A stratified random sample
- (c) A cluster sample
- (d) A systematic random sample
- (e) A voluntary response sample

37. Your statistics class has 30 students. You want to ask an SRS of 5 students from your class whether they use a mobile device for the online quizzes. 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

- (a) 14, 45, 92, 60, 56.
- (b) 14, 31, 03, 10, 22.
- (c) 14, 03, 10, 22, 22.
- (d) 14, 03, 10, 22, 06.
- (e) 14, 03, 10, 22, 11.

38. Suppose that 35% of the voters in a state are registered as Republicans, 40% as Democrats, and 25% as Independents. A newspaper wants to select a sample of 1000 registered voters to predict the outcome of the next election. If it randomly selects 350 Republicans,

randomly selects 400 Democrats, and randomly selects 250 Independents, did this sampling procedure result in a simple random sample of registered voters from this state?

- (a) Yes, because each registered voter had the same chance of being chosen.
- (b) Yes, because random chance was involved.
- (c) No, because not all registered voters had the same chance of being chosen.
- (d) No, because a different number of registered voters was selected from each party.
- (e) No, because not all possible groups of 1000 registered voters had the same chance of being chosen.

39. A local news agency conducted a survey about unemployment by randomly dialing phone numbers during the work day until it gathered responses from 1000 adults in its state. In the survey, 19% of those who responded said they were not currently employed. In reality, only 6% of the adults in the state were not currently employed at the time of the survey. Which of the following best explains the difference in the two percentages?

- (a) The difference is due to sampling variability. We shouldn't expect the results of a random sample to match the truth about the population every time.
- (b) The difference is due to response bias. Adults who are employed are likely to lie and say that they are unemployed.
- (c) The difference is due to undercoverage bias. The survey included only adults and did not include teenagers who are eligible to work.
- (d) The difference is due to nonresponse bias. Adults who are employed are less likely to be available for the sample than adults who are unemployed.
- (e) The difference is due to voluntary response. Adults are able to volunteer as a member of the sample.

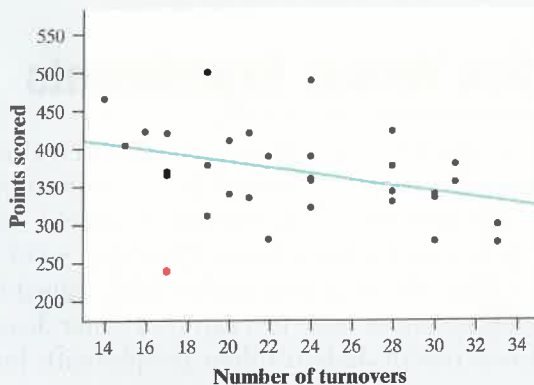
40. A simple random sample of 1200 adult Americans is selected, and each person is asked the following question: “In light of the huge national deficit, should the government at this time spend additional money to send humans to Mars?” Only 39% of those responding answered “Yes.” This survey

- (a) is reasonably accurate because it used a large simple random sample.
- (b) needs to be larger because only about 24 people were drawn from each state.
- (c) probably understates the percent of people who favor sending humans to Mars.

- (d) is very inaccurate but neither understates nor overstates the percent of people who favor sending humans to Mars. Because simple random sampling was used, it is unbiased.
- (e) probably overstates the percent of people who favor sending humans to Mars.

Recycle and Review

41. **Don't turn it over (3.2)** How many points do turnovers cost teams in the NFL? The scatterplot shows the relationship between x = number of turnovers and y = number of points scored by teams in the NFL during 2015, along with the least-squares regression line $\hat{y} = 460.2 - 4.084x$.



- (a) Interpret the slope of the regression line in context.
- (b) For this regression line, $s = 57.3$. Interpret this value.
- (c) Calculate and interpret the residual for the San Francisco 49ers, who had 17 turnovers and scored 238 points.
- (d) How does the point for the 49ers affect the least-squares regression line and standard deviation of the residuals? Explain your answer.
42. **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 to calculate bandwidth is to find the 95th percentile of a company's usage based on samples of hundreds of 5-minute intervals during a month.
- (a) Explain what "95th percentile" means in this setting.
- (b) Is it possible to determine the z-score for a usage total that is at the 95th percentile? If so, find the z-score. If not, explain why not.

SECTION 4.2 Experiments

LEARNING TARGETS *By the end of the section, you should be able to:*

- Explain the concept of confounding and how it limits the ability to make cause-and-effect conclusions.
- Distinguish between an observational study and an experiment, and identify the explanatory and response variables in each type of study.
- Identify the experimental units and treatments in an experiment.
- Describe the placebo effect and the purpose of blinding in an experiment.
- Describe how to randomly assign treatments in an experiment using slips of paper, technology, or a table of random digits.
- Explain the purpose of comparison, random assignment, control, and replication in an experiment.
- Describe a completely randomized design for an experiment.
- Describe a randomized block design and a matched pairs design for an experiment and explain the purpose of blocking in an experiment.

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 record the behavior of animals in the wild or

track the medical history of volunteers to look for associations between variables such as type of diet, amount of exercise, and blood pressure.

DEFINITION Observational study

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

Section 4.2 is about statistical designs for experiments, a very different way to produce data.

Observational Studies Versus Experiments

Is taking a vitamin D supplement good for you? Hundreds of observational studies have looked at the relationship between vitamin D concentration in a person's blood and various health outcomes.¹⁸ In one observational study, researchers found that teenage girls with higher vitamin D intakes were less likely to suffer broken bones.¹⁹ Other observational studies have shown that people with higher vitamin D concentration have less cardiovascular disease, better cognitive function, and less risk of diabetes than people with lower concentrations of vitamin D.

In the observational studies involving vitamin D and diabetes, the **explanatory variable** is vitamin D concentration in the blood and the **response variable** is diabetes status—whether or not the person developed diabetes.

DEFINITION Response variable, Explanatory variable

A **response variable** measures an outcome of a study. An **explanatory variable** may help explain or predict changes in a response variable.

Unfortunately, it is very difficult to show that taking vitamin D *causes* a lower risk of diabetes using an observational study. As shown in the table, there are many possible differences between the group of people with high vitamin D concentration and the group of people with low vitamin D concentration. Any of these differences could be causing the difference in diabetes risk between the two groups of people.

Variable	Group 1	Group 2
Vitamin D concentration (explanatory)	High vitamin D concentration	Low vitamin D concentration
Quality of diet	Better diet	Worse diet
Amount of exercise	More exercise	Less exercise
⋮	⋮	⋮
Amount of vitamin supplementation	More likely to take other vitamins	Less likely to take other vitamins
Diabetes status (response)	Less likely to have diabetes	More likely to have diabetes



Some people call a variable that results in confounding, like diet in this case, a *confounding variable*.

For example, it is possible that people who have healthier diets eat lots of foods that are high in vitamin D. Likewise, it is possible that people with healthier diets are less likely to develop diabetes. Vitamin D concentration may not have anything to do with diabetes status, even though there is an association between the two variables. In this case, we say there is **confounding** between vitamin D concentration and diet because we cannot tell which variable is causing the change in diabetes status.

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) is associated with the response 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.

Likewise, because sun exposure increases vitamin D concentration, it is possible that people who exercise a lot outside have higher concentrations of vitamin D. If people who exercise a lot are also less likely to get diabetes, then amount of exercise and vitamin D concentration are confounded—we can't say which variable is the cause of the smaller diabetes risk.

EXAMPLE**Smoking and ADHD**
Confounding

PROBLEM: In a study of more than 4700 children, researchers from Cincinnati Children's Hospital Medical Center found that those children whose mothers smoked during pregnancy were more than twice as likely to develop ADHD as children whose mothers had not smoked.²⁰ Explain how confounding makes it unreasonable to conclude that a mother's smoking during pregnancy causes an increase in the risk of ADHD in her children based on this study.

SOLUTION:

It is possible that the mothers who smoked during pregnancy were also more likely to have unhealthy diets. If people with unhealthy diets are also more likely to have children with ADHD, then it could be that unhealthy diets caused the increase in ADHD risk, not smoking.



Al/Getty Images

Notice that the solution describes how diet might be associated with the explanatory variable (smoking status) and with the response variable (ADHD status).

FOR PRACTICE, TRY EXERCISE 43

Observational studies of the effect of an explanatory variable on a response variable often fail because of confounding between the explanatory variable and one or more other variables. In contrast to observational studies, **experiments** don't just observe individuals or ask them questions. They actively impose some *treatment* 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 Experiment

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

To determine if taking vitamin D actually causes a reduction in diabetes risk, researchers in Norway performed an experiment. The researchers randomly assigned 500 people with pre-diabetes to either take a high dose of vitamin D or to take a **placebo**—a pill that looked exactly like the vitamin D supplement but contained no active ingredient. After 5 years, about 40% of the people in each group were diagnosed with diabetes.²¹ In other words, the association between vitamin D concentration and diabetes status disappeared when comparing two groups that were roughly the same to begin with.

DEFINITION Placebo

A **placebo** is a treatment that has no active ingredient, but is otherwise like other treatments.

The experiment in Norway avoided confounding by letting chance decide who took vitamin D and who didn't. That way, people with healthier diets were split about evenly between the two groups. So were people who exercise a lot and people who take other vitamins. *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

Facebook and financial incentives Observational studies and experiments

PROBLEM: In each of the following settings, identify the explanatory and response variables. Then determine if each is an experiment or an observational study. Explain your reasoning.

- (a) In a study conducted by researchers at the University of Texas, people were asked about their social media use and satisfaction with their marriage. Of the heavy social media users, 32% had thought seriously about leaving their spouse. Only 16% of non-social media users had thought seriously about leaving their spouse.²²
- (b) In a diet study using 100 overweight volunteers, 50 volunteers were randomly assigned to receive weight-loss counseling, monthly weigh-ins, and a three-month gym pass. The other 50 volunteers were given financial incentives (earning \$20 for losing 4 pounds in a month or paying \$20 otherwise) along with the counseling, weigh-ins, and gym pass. The group with the financial incentives lost 6.7 more pounds, on average.²³



Postislav_Sedlacek/Shutterstock.com

SOLUTION:

- (a) Explanatory variable: Frequency of social media use. Response variable: Marital satisfaction. This is an *observational study* because people weren't assigned to use social media or not.
- (b) Explanatory variable: Whether or not financial incentives were given. Response variable: Amount of weight lost. This is an *experiment* because researchers gave some volunteers financial incentives and did not give financial incentives to the other volunteers.

In part (a), the response variable is not the *percent* who thought about leaving their spouse. This percent is a summary of all the responses. Likewise, in part (b), the response variable is not the *average* weight loss. This average is a summary of all the responses.

FOR PRACTICE, TRY EXERCISE 45

In part (a) of the example, it would be incorrect to conclude that using social media causes marital dissatisfaction. It could be that other variables are confounded with social media use—or even that marital dissatisfaction is causing increased social media use. In part (b), it is reasonable to conclude that the financial incentives caused the increase in weight loss because this was a well-designed experiment.



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 2–4 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."²⁴ 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 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** or **subjects**) to observe the response.

DEFINITION Treatment, Experimental unit, 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. An **experimental unit** is the object to which a treatment is randomly assigned. When the experimental units are human beings, they are often called **subjects**.

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

EXAMPLE**How can we prevent malaria?****Vocabulary of experiments**

PROBLEM: Malaria causes hundreds of thousands of deaths each year, with many of the victims being children. Will regularly screening children for the malaria parasite and treating those who test positive reduce the proportion of children who develop the disease? Researchers worked with children in 101 schools in Kenya, randomly assigning half of the schools to receive regular screenings and follow-up treatments and the remaining schools to receive no regular screening. Children at all 101 schools were tested for malaria at the end of the study.²⁵ Identify the treatments and the experimental units in this experiment.

SOLUTION:

This experiment compares two treatments: (1) regular screenings and follow-up treatments and (2) no regular screening. The experimental units are 101 schools in Kenya.

Note that the experimental units are the schools, not the students. The decision about who to screen was made school by school, not student by student. All students at the same school received the same treatment.

FOR PRACTICE, TRY EXERCISE 49

In the malaria experiment, there was one explanatory variable: screening status. In other experiments, there are multiple explanatory variables. Sometimes, these explanatory variables are called **factors**. In an experiment with multiple factors, the treatments are formed using the various levels of each of the factors.

DEFINITION Factor, Levels

In an experiment, a **factor** is a variable that is manipulated and may cause a change in the response variable. The different values of a factor are called **levels**.

Here's an example of a multifactor experiment.

EXAMPLE**The five-second rule****Experiments with multiple explanatory variables**

PROBLEM: Have you ever dropped a tasty piece of food on the ground, then quickly picked it up and eaten it? If so, you probably thought about the “five-second rule,” which states that a piece of food is safe to eat if it has been on the floor less than 5 seconds. The rule is based on the belief that bacteria need time to transfer from the floor to the food. But does it work?

Researchers from Rutgers University put the five-second rule to the test. They used four different types of food: watermelon, bread, bread with butter, and gummy candy. They dropped the food onto four different surfaces: stainless steel, ceramic tile, wood, and carpet. And they waited for four different lengths of time: less than 1 second, 5 seconds, 30 seconds, and 300 seconds. Finally, they used bacteria prepared two different ways: in a tryptic soy broth or peptone buffer. Once the bacteria were ready, the researchers spread them out on the different surfaces and started dropping food.²⁶



Zaia Snively

- (a) List the factors in this experiment and the number of levels for each factor.
- (b) If the researchers used every possible combination to form the treatments, how many treatments were included in the experiment?
- (c) List two of the treatments.

SOLUTION:

- (a) *Type of food (4 levels), type of surface (4 levels), amount of time (4 levels), method of bacterial preparation (2 levels)*
- (b) $4 \times 4 \times 4 \times 2 = 128$ different treatments
- (c) *Watermelon/stainless steel/less than 1 second/tryptic soy broth; gummy candy/wood/300 seconds/peptone buffer*

FOR PRACTICE, TRY EXERCISE 51

What did the researchers discover? The wetter foods had greater bacterial transfer and food dropped on carpet had the least bacterial transfer. There was greater bacterial transfer the longer the food was on the surface, although there was some transfer that happened almost instantaneously. Overall, the researchers concluded that the type of food and type of surface were at least as important as the amount of time the food remained on the surface.

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. For example, although longer time was associated with more bacterial transfer in general, this relationship might not be true for very moist food.

Designing Experiments: Comparison

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.

Does caffeine affect pulse rate? Many students regularly consume caffeine to help them stay alert. So it seems plausible that taking caffeine might increase an individual's pulse rate. Is this true? One way to investigate this claim is to ask volunteers to measure their pulse rates, drink some cola with caffeine, measure their pulse rates again after 10 minutes, and calculate the increase in pulse rate.

This experiment has a very simple design. A group of subjects (the students) were exposed to a treatment (the cola with caffeine), and the outcome (change in pulse rate) was observed. Here is the design:

Students → Cola with caffeine → Change in pulse rate

Unfortunately, even if the pulse rate of every student went up, we couldn't attribute the increase to caffeine. Perhaps the excitement of being in an experiment made their pulse rates increase. Maybe it was the sugar in the cola and not the caffeine. Perhaps their teacher told them a funny joke during the 10-minute



Zaia Snively

waiting period and made everyone laugh. In other words, there are many other variables that are potentially confounded with taking caffeine.

Many laboratory experiments use a design like the one in the caffeine 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 varied conditions. *Outside the lab, badly designed experiments often yield worthless results because of confounding.*

The remedy for the confounding in the caffeine example is to do a comparative experiment with two groups: one group that receives caffeine and a **control group** that does not receive caffeine.

DEFINITION Control group

In an experiment, a **control group** is used to provide a baseline for comparing the effects of other treatments. Depending on the purpose of the experiment, a control group may be given an inactive treatment (placebo), an active treatment, or no treatment at all.

In all other aspects, these groups should be treated exactly the same so that the only difference is the caffeine. That way, if there is convincing evidence of a difference in the average increase in pulse rates, we can safely conclude it was *caused* by the caffeine. This means that one group could get regular cola with caffeine, while the control group gets caffeine-free cola. Both groups would get the same amount of sugar, so sugar consumption would no longer be confounded with caffeine intake. Likewise, both groups would experience the same events during the experiment, so what happens during the experiment won't be confounded with caffeine intake either.

EXAMPLE

Preventing malaria Control groups

PROBLEM: In an earlier example, we described an experiment in which researchers randomly assigned 101 schools in Kenya to either receive regular screenings and follow-up treatments or to receive no regular screenings. Explain why it was necessary to include a control group of schools that didn't receive regular screenings.

SOLUTION:

The purpose of the control group is to provide a baseline for comparing the effect of the regular screenings and follow-up treatments. Otherwise, researchers wouldn't be able to determine if a decrease in malaria rates was due to the treatment or some other change that occurred during the experiment (like a drought that killed off mosquitoes, slowing the spread of malaria).



Alexander Joe/Getty Images

FOR PRACTICE, TRY EXERCISE 55

A control group was essential in the malaria experiment to determine if screening was effective. However, not all experiments include a control group—as long as comparison takes place. In the experiment about the five-second rule, there were 128 different treatments being compared and no control group. A control group wasn't essential in this experiment because researchers were interested in comparing different amounts of time on the floor, different types of food, and different types of surfaces.

Designing Experiments: Blinding and the Placebo Effect

In the caffeine experiment, we used comparison to help prevent confounding. But even when there is comparison, confounding is still possible. If the subjects in the experiment know what type of soda they are receiving, the expectations of the two groups will be different. The knowledge that a subject is receiving caffeine may increase his or her pulse rate, apart from the caffeine itself. This is an example of the **placebo effect**.

DEFINITION Placebo effect

The **placebo effect** describes the fact that some subjects in an experiment will respond favorably to any treatment, even an inactive treatment.

In one 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.²⁷ When the ailment is psychological, like depression, some experts think that the placebo effect accounts for about three-quarters of the effect of the most widely used drugs.²⁸

Because of the placebo effect, it is important that subjects don't know what treatment they are receiving. It is also better if the people interacting with the subjects and measuring the response variable don't know which subjects are receiving which treatment. When neither group knows who is receiving which treatment, the experiment is **double-blind**. Other experiments are **single-blind**.

DEFINITION Double-blind, Single-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.

In a **single-blind** experiment, either the subjects don't know which treatment they are receiving or the people who interact with them and measure the response variable don't know which subjects are receiving which treatment.

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. For example, 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.

EXAMPLE

Do magnets repel pain? Blinding and the placebo effect

PROBLEM: 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 if this type of therapy really worked. They designed a double-blind experiment to find out. A total of 50 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 at random from a box with a mixture of active and inactive magnets. The chosen magnet was applied to the site of the pain for 45 minutes. After being treated, each patient was again asked to rate the level of pain from 0 to 10.²⁹



Eric O'Connell/Getty Images

- (a) Explain what it means for this experiment to be double-blind.
- (b) Why was it important for this experiment to be double-blind?

SOLUTION:

- (a) Neither the subjects nor the doctors applying the magnets and recording the pain ratings knew which subjects had the active magnets and which had the inactive magnets.
- (b) If subjects knew they were receiving an active treatment, researchers wouldn't know if any improvement was due to the magnets or to the expectation of getting better (the placebo effect). If the doctors knew which subjects received which treatments, they might treat one group of subjects differently from the other group. This would make it difficult to know if the magnets were the cause of any improvement.

FOR PRACTICE, TRY EXERCISE 59



CHECK YOUR UNDERSTANDING

A new analysis is casting doubt on a claimed benefit of omega-3 fish oil. For years, doctors have been recommending eating fish and taking fish oil supplements to prevent heart disease. But the new analysis reviewed 20 previous studies and showed that the effects of omega-3 aren't as great as once suspected. One reason is that an early trial of omega-3 supplements was conducted as an open-label study.³⁰ In this type of study, both patients and researchers know who is receiving which treatment.

1. Describe a potential problem with an open-label study in this context.
2. Describe how you can fix the problem identified in Question 1.

Designing Experiments: Random Assignment

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, confounding will result. If we allow students to choose what type of cola they will drink in the caffeine experiment, students who consume caffeine on a regular basis might be more likely to choose the regular cola. Due to their caffeine tolerance, these students' pulse rates might not increase as much as other students' pulse rates. In this case, caffeine tolerance would be confounded with the amount of caffeine consumed, making it impossible to conclude cause and effect.

To create roughly equivalent groups at the beginning of an experiment, we use **random assignment** to determine which experimental units get which treatment.

DEFINITION Random assignment

In an experiment, **random assignment** means that experimental units are assigned to treatments using a chance process.

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

EXAMPLE

Caffeine and pulse rates How random assignment works

PROBLEM: A total of 20 students have agreed to participate in an experiment comparing the effects of caffeinated cola and caffeine-free cola on pulse rates. Describe how you would randomly assign 10 students to each of the two treatments:

- (a) Using 20 identical slips of paper
- (b) Using technology
- (c) Using Table D

SOLUTION:

- (a) On 10 slips of paper, write the letter "A"; on the remaining 10 slips, write the letter "B." Shuffle the slips of paper and hand out one slip of paper to each volunteer. Students who get an "A" slip receive the cola with caffeine and students who get a "B" slip receive the cola without caffeine.
- (b) Label each student with a different integer from 1 to 20. Then randomly generate 10 different integers from 1 to 20. The students with these labels receive the cola with caffeine. The remaining 10 students receive the cola without caffeine.
- (c) Label each student with a different integer from 01 to 20. Go to a line of Table D and read two-digit groups moving from left to right. The first 10 different labels between 01 and 20 identify the 10 students who receive cola with caffeine. The remaining 10 students receive the caffeine-free cola. Ignore groups of digits from 21 to 00.



Zaia Snively

When describing a method of random assignment, don't stop after creating the groups. Make sure to identify which group gets which treatment.

When using a random number generator or a table of random digits to assign treatments, make sure to account for the possibility of repeated numbers when describing your method.

Random assignment should distribute the students who regularly consume caffeine in roughly equal numbers to each group. It should also balance out the students with high metabolism and those with larger body sizes in the caffeine and caffeine-free groups. Random assignment helps ensure that the effects of other variables (e.g., caffeine tolerance, metabolism, or body size) are spread evenly among the two groups.

Designing Experiments: Control

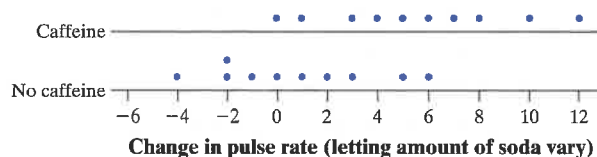
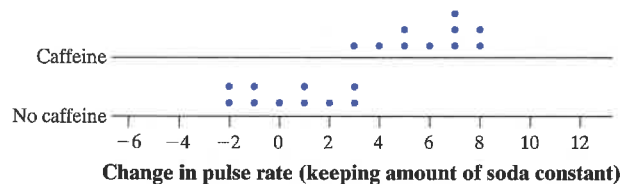
Although random assignment should create two groups of students that are roughly equivalent to begin with, we still have to ensure that the only consistent difference between the groups during the experiment is the type of cola they receive. We can **control** the effects of some variables by keeping them the same for both groups. For example, we should make both treatments contain the same amount of sugar. If one group got regular cola and the other group got caffeine-free *diet* cola, then the amount of sugar would be confounded with the amount of caffeine—we wouldn't know if it was the sugar or the caffeine that was causing a change in pulse rates.

DEFINITION Control

In an experiment, **control** means keeping other variables constant for all experimental units.

We also want to control other variables to reduce the variability in the response variable. Suppose we let volunteers in both groups choose how much cola they want to drink. In that case, the changes in pulse rate would be more variable than if we made sure each subject drank the same amount of soda. Letting the amount of cola vary will make it harder to determine if caffeine is really having an effect on pulse rates.

The dotplots on the left show the results of an experiment in which the amount of cola consumed was the same for all participating students. Because there is so little overlap in these graphs, it seems clear that caffeine increases pulse rates. The dotplots on the right show the results of an experiment in which the students were permitted to choose how much or how little cola they consumed. Notice that the centers of the distributions haven't changed, but the distributions are much more variable. The increased overlap in the graphs makes the evidence supporting the effect of caffeine less convincing.



After randomly assigning treatments and controlling other variables, the two groups should be about the same, except for the treatments. Then a difference in the average change in pulse rate must be due either to the treatments themselves—or to the random assignment. We can't say that *any* difference

between the average pulse rate changes for students in the two groups must be caused by the difference in caffeine. There would be *some* difference, even if both groups received the same type of cola, because the random assignment is unlikely to produce two groups that are exactly equivalent with respect to every variable that might affect pulse rate.

Designing Experiments: Replication

Would you trust an experiment with just one student in each group? No, because the results would depend too much on which student was assigned to the caffeinated cola. However, if we randomly assign many subjects to each group, the effects of chance will balance out, and there will be little difference in the average responses in the two groups—unless the treatments themselves cause a difference. This is the idea of **replication**.

DEFINITION Replication

In an experiment, **replication** means using enough experimental units to distinguish a difference in the effects of the treatments from chance variation due to the random assignment.



In statistics, replication means “use enough subjects.” In other fields, the term *replication* has a different meaning. In these fields, replication means conducting an experiment in one setting and then having other investigators conduct a similar experiment in a different setting. That is, replication means repeatability.

Experiments: Putting It All Together

The following box summarizes the four key principles of experimental design: comparison, random assignment, control, and replication.

PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles for designing experiments are as follows:

1. **Comparison.** Use a design that compares two or more treatments.
2. **Random assignment.** Use chance to assign experimental units to treatments. Doing so helps create roughly equivalent groups of experimental units by balancing the effects of other variables among the treatment groups.
3. **Control.** Keep other variables the same for all groups, especially variables that are likely to affect the response variable. Control helps avoid confounding and reduces variability in the response variable.
4. **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 Principles of experimental design

PROBLEM: 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 if 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 the levels of these factors form the four treatments shown in the diagram. 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 placebo that looked and tasted like the aspirin but had no active ingredient. 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.³¹ It did not appear, however, that beta-carotene had any effect on preventing cancer.



SCIENCE PHOTO LIBRARY/AGE Fotostock

		Factor 2: Beta-carotene	
		Yes	No
Factor 1: Aspirin	Yes	  Aspirin Beta-carotene	  Aspirin Placebo
	No	  Placebo Beta-carotene	  Placebo Placebo

- Explain how this experiment used comparison.
- Explain the purpose of randomly assigning the physicians to the four treatments.
- Name two variables that were controlled in this experiment and why it was beneficial to control these variables.
- Explain how this experiment used replication. What is the purpose of replication in this context?

SOLUTION:

- Researchers used a design that compared each of the active treatments to a placebo.
- Random assignment helped ensure that the four groups of physicians were roughly equivalent at the beginning of the experiment.
- The experiment used subjects of the same gender and same occupation. Using only male physicians helps to reduce the variability in the response variables.
- There were over 5000 subjects in each treatment group. This helped ensure that the difference in heart attacks was due to the aspirin and not to chance variation in the random assignment.

If women and people with other occupations were included, the results might be more variable, making it harder to determine the effects of aspirin and beta-carotene. However, using only male physicians means we don't know how females or other males would respond to these treatments.

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."³² Doctors are supposed to know what this means. Now you know, too.

The Physicians' Health Study shows how well-designed experiments can yield good evidence that differences in the treatments cause the differences we observe in the response.



CHECK YOUR UNDERSTANDING

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 using 60 households to compare these two approaches (display, chart) with a group of customers who receive information about energy consumption but no help in monitoring electricity use.

1. Explain why it was important to have a control group that didn't get the display or the chart.
2. Describe how to randomly assign the treatments to the 60 households.
3. What is the purpose of randomly assigning treatments in this context?

Completely Randomized Designs

The diagram in Figure 4.1 presents the details of the caffeine experiment: random assignment, the sizes of the groups and which treatment they receive, and the response variable. This type of design is called a **completely randomized design**.

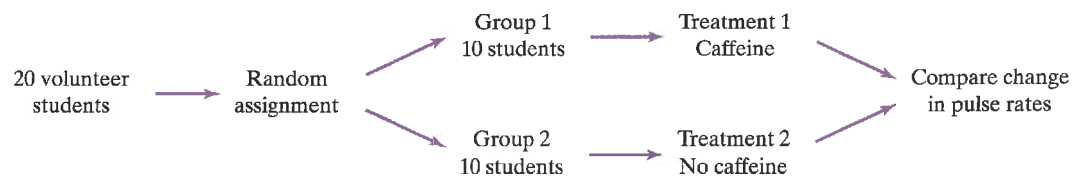


FIGURE 4.1 Outline of a completely randomized design to compare caffeine and no caffeine.

DEFINITION Completely randomized design

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

Although there are good statistical reasons for using treatment groups that are about equal in size, 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.

EXAMPLE**Chocolate milk and concussions**
Completely randomized design

Ann Heath

PROBLEM: “Concussion-Related Measures Improved in High School Football Players Who Drank New Chocolate Milk” announced a recent headline.³³ In the study, researchers compared a group of concussed football players given a new type of chocolate milk with a group of concussed football players who received no treatment.

- Explain why it isn’t reasonable to conclude that the new type of chocolate milk is effective for treating high school football players with concussions based on this study.
- To test the effectiveness of the new type of chocolate milk, you recruit 50 high school football players who suffered a concussion in the previous 24 hours to participate in an experiment. Write a few sentences describing a completely randomized design for this experiment.

SOLUTION:

- It is possible that the group who received the new type of chocolate milk improved because they knew they were being treated and expected to get better, not because of the new chocolate milk.
- Number the players from 1 to 50. Use a random number generator to produce 25 different integers from 1 to 50 and give the new type of chocolate milk to the players with these numbers. Give regular chocolate milk to the remaining 25 players. Compare the concussion-related measures for the two groups.

FOR PRACTICE, TRY EXERCISE 69**AP® EXAM TIP**

If you are asked to describe a completely randomized design, stay away from flipping coins. For example, suppose we ask each student in the caffeine experiment to toss a coin. If it’s heads, then the student will drink the cola with caffeine. If it’s tails, then the student will drink the caffeine-free cola. As long as all 20 students toss a coin, this is still a completely randomized design. Of course, the two groups are unlikely to contain exactly 10 students because it is unlikely that 20 coin tosses will result in a perfect 50-50 split between heads and tails.

The problem arises if we try to force the two groups to have equal sizes. Suppose we continue to have students toss coins until one of the groups has 10 students and then place the remaining students in the other group. In this case, the last two students in line are very likely to end up in the same group. However, in a completely randomized design, the last two subjects should only have a 50% chance of ending up in the same group.

Randomized Block Designs

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of comparison, random assignment, control, 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.

Suppose that a mobile phone company is considering two different keyboard designs (A and B) for its new smartphone. The company decides to perform an experiment to compare the two keyboards using a group of 10 volunteers. The response variable is typing speed, measured in words per minute.



How should the company address the fact that four of the volunteers already use a smartphone, whereas the remaining six volunteers do not? They could use a completely randomized design and hope that the random assignment distributes the smartphone users and non-smartphone users about evenly between the group using keyboard A and the group using keyboard B. Even so, there might be a lot of variability in typing speed within both treatment groups because some members of each treatment group are more familiar with smartphones than the others. This additional variability might make it difficult to detect a difference in the effectiveness of the two keyboards. What should the researchers do?

Because the company knows that experience with smartphones will affect typing speed, they could start by separating the volunteers into two groups—one with experienced smartphone users and one with inexperienced smartphone users. Each of these groups of similar subjects is known as a **block**. Within each block, the company could then randomly assign half of the subjects to use keyboard A and the other half to use keyboard B. To control other variables, each subject should be given the same passage to type while in a quiet room with no distractions. This **randomized block design** helps account for the variation in typing speed that is due to experience with smartphones.

DEFINITION Block, 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.

Figure 4.2 outlines the randomized block design for the smartphone experiment. The subjects are first separated into blocks based on their experience with smartphones. Then the two treatments are randomly assigned within each block.

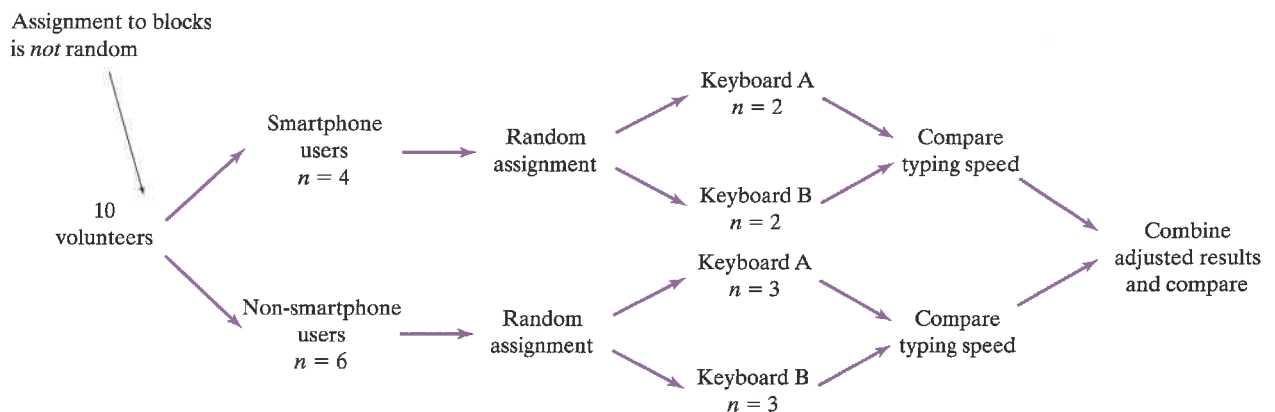
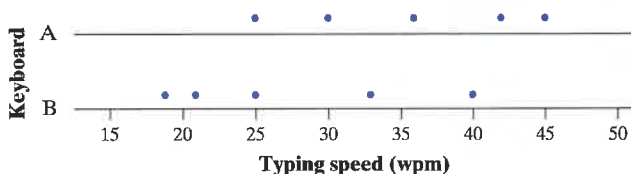


FIGURE 4.2 Outline of a randomized block design for the smartphone experiment. The blocks consist of volunteers who have used smartphones and volunteers who have not used smartphones. The treatments are keyboard A and keyboard B.

Using a randomized block design allows us to account for the variation in the response that is due to the blocking variable of smartphone experience. This makes it easier to determine if one treatment is really more effective than the other.

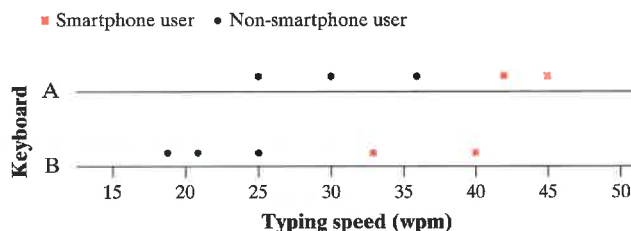
To see how blocking helps, let's look at the results of the smartphone experiment. In the block of 4 smartphone users, 2 were randomly assigned to use keyboard A and the other 2 were assigned to use keyboard B. Likewise, in the block of 6 non-smartphone users, 3 were randomly assigned to use keyboard A and the

other 3 were assigned to use keyboard B. Each of the 10 volunteers typed the same passage and the typing speed was recorded. Here are the results:

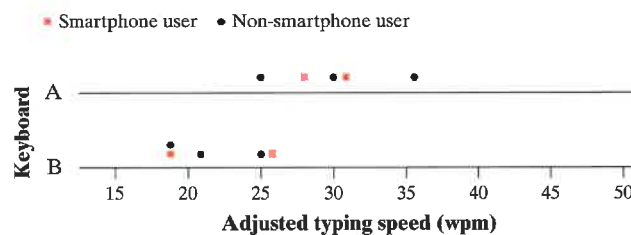


There is some evidence that keyboard A results in higher typing speeds, but the evidence isn't that convincing. Enough overlap occurs in the two distributions that the differences might simply be due to the chance variation in the random assignment.

If we compare the results for the two keyboards within each block, however, a different story emerges. Among the 4 smartphone users (indicated by the red squares), keyboard A was the clear winner. Likewise, among the 6 non-smartphone users (indicated by the black dots), keyboard A was also the clear winner.



The overlap in the first set of dotplots was due almost entirely to the variation in smartphone experience—smartphone users were generally faster than non-smartphone users, regardless of which keyboard they used. In fact, the average typing speed for the smartphone users was 40, while the average typing speed for non-smartphone users was only 26, a difference of 14 words per minute. To account for the variation created by the difference in smartphone experience, let's subtract 14 from each of the typing speeds in the block of smartphone users to “even the playing field.” Here are the results:



Because we accounted for the variation due to the difference in smartphone experience, the variation in each of the distributions has been reduced. There is now much less overlap between the two distributions, meaning that the evidence in favor of keyboard A is much more convincing. *When blocks are formed wisely, it is easier to find convincing evidence that one treatment is more effective than another.*

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 because some subjects dropped out during the study.”

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 variation among the experimental units. In other words, the experimenter will form blocks using the variables that are the best predictors of the response variable. Random assignment will then average out the effects of the remaining other variables and allow a fair 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

Should I use the popcorn button? Blocking in an experiment

PROBLEM: A popcorn lover wants to determine if it is better to use the “popcorn button” on her microwave oven or use the amount of time recommended on the bag of popcorn. To measure how well each method works, she will count the number of unpopped kernels remaining after popping. To obtain the experimental units, she goes to the store and buys 10 bags each of 4 different varieties of microwave popcorn (butter, cheese, natural, and kettle corn), for a total of 40 bags.



Ann Hebl

- Describe a randomized block design for this experiment. Justify your choice of blocks.
- Explain why a randomized block design might be preferable to a completely randomized design for this experiment.

SOLUTION:

- Form blocks based on variety, because the number of unpopped kernels is likely to differ by variety. Randomly assign 5 bags of each variety to the popcorn button treatment and 5 to the timed treatment by placing all 10 bags of a particular variety in a large box. Shake the box, pick 5 bags without looking, and assign them to be popped using the popcorn button. The remaining 5 bags will be popped using the instructions on the bags. Repeat this process for the remaining 3 varieties. After popping each of the 40 bags in random order, count the number of unpopped kernels in each bag and compare the results within each variety. Then combine the results from the 4 varieties after accounting for the difference in average response for each variety.
- A randomized block design accounts for the variability in the number of unpopped kernels created by the different varieties of popcorn (butter, cheese, natural, kettle). This makes it easier to determine if using the microwave button is more effective for reducing the number of unpopped kernels.

It is important to pop the bags in random order so that changes over time (e.g., temperature, humidity) aren't confounded with the explanatory variable. For example, if the 20 “popcorn button” bags are popped last when the room temperature is greater, we wouldn't know if using the popcorn button or the warmer temperature was the cause of a difference in the number of unpopped kernels.

FOR PRACTICE, TRY EXERCISE 71

Another way to address the variability in unpopped kernels created by the different varieties is to use only one variety of popcorn in the experiment. Because variety of popcorn is no longer a variable, it will not be a source of variability. Of course, this means that the results of the experiment only apply to that one variety of popcorn—not ideal for the popcorn lover in the example!

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. The random assignment of subjects to treatments is done within each matched pair. Just as with other forms of blocking, matching helps account for the variation due to the variable(s) used to form the pairs.

DEFINITION Matched pairs design

A **matched pairs design** is a common experimental design for comparing two treatments that uses blocks of size 2. In some matched pairs designs, two very similar experimental units are paired and the two treatments are randomly assigned within each pair. In others, each experimental unit receives both treatments in a random order.



J-Elgaard/Getty Images

Suppose we want to investigate if listening to classical music while taking a math test affects performance. A total of 30 students in a math class volunteer to take part in the experiment. The difference in mathematical ability among the volunteers is likely to create additional variation in the test scores, making it harder to see the effect of classical music. To account for this variation, we could pair the students by their grade in the class—the two students with the highest grades are paired together, the two students with the next highest grades are paired together, and so on. Within each pair, one student is randomly assigned to take a math test while listening to classical music and the other member of the pair is assigned to take the math test in silence.

Sometimes, each “pair” in a matched pairs design consists of just one experimental unit that gets both treatments in random order. In the experiment about the effect of listening to classical music, we could have each student take a math test in both conditions. To decide the order, we might flip a coin for each student. If the coin lands on heads, the student takes a math test with classical music playing today and a similar math test without music playing tomorrow. If it lands on tails, the student does the opposite—no music today and classical music tomorrow.

Randomizing the order of treatments is important to avoid confounding. Suppose everyone did the classical music treatment on the first day and the no-music treatment on the second day, but the air conditioner wasn’t working on the second day. We wouldn’t know if any difference in mean test score was due to the difference in treatment or the difference in room temperature.

EXAMPLE

Will an additive improve my mileage?

Matched pairs design

PROBLEM: A consumer organization wants to know if using a certain fuel additive increases the fuel efficiency (in miles per gallon, or mpg) of cars. A total of 20 cars of different types are available for testing. Design an experiment that uses a matched pairs design to investigate this question. Explain your method of pairing.

SOLUTION:

Give each car both treatments. It is reasonable to think that some cars are more fuel efficient than others, so using each car as its own “pair” accounts for the variation in fuel efficiency in the experimental units. For each car, randomly assign the order in which the treatments are assigned by flipping a coin. Heads indicates using the additive first and no additive second. Tails indicates using no additive first and then the additive second. For each car, record the fuel efficiency (mpg) after using each treatment.

FOR PRACTICE, TRY EXERCISE 77

In the preceding example, it is also possible to form pairs of two similar cars. For instance, we could pair together the two most fuel-efficient cars, the next two most fuel-efficient cars, and so on. This is less ideal, however, because there will

still be some differences between the members of each pair that may cause additional variation in the results. Using the same car twice creates perfectly matched “pairs,” and it also doubles the number of pairs used in the experiment. Both these features make it easier to find convincing evidence that the gas additive is effective, if it really is effective.



CHECK YOUR UNDERSTANDING

Researchers would like to design an experiment to compare the effectiveness of three different advertisements for a new television series featuring the works of Jane Austen. There are 300 volunteers available for the experiment.

1. Describe a completely randomized design to compare the effectiveness of the three advertisements.
2. Describe a randomized block design for this experiment. Justify your choice of blocks.
3. Why might a randomized block design be preferable in this context?

Section 4.2

Summary

- 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 variable can't be distinguished from each other.
- We can produce data to answer specific questions using **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. Experiments are the best way to show cause and effect.
- 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 the **levels** of the explanatory variables (also called **factors**).
- Some experiments give a **placebo** (fake treatment) to a **control group**. That helps prevent confounding due to the **placebo effect**, whereby some patients get better because they expect the treatment to work.
- 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 group knows and the other doesn't, then the experiment is **single-blind**.
- The basic principles of experimental design are:
 - **Comparison:** Use a design that compares two or more treatments.
 - **Random assignment:** Use chance (slips of paper, a random number generator, a table of random digits) to assign experimental units to treatments. This helps create roughly equivalent groups before treatments are imposed.
 - **Control:** Keep other variables the same for all groups. Control helps avoid confounding and reduces the variation in responses, making it easier to decide if a treatment is effective.

▪ **Replication:** Impose each treatment on enough experimental units so that the effects of the treatments can be distinguished from chance differences between the groups.

- In a **completely randomized design**, the experimental units are assigned to the treatments completely by chance.
- A **randomized block design** forms groups (**blocks**) of experimental units that are similar with respect to a variable that is expected to affect the response. Treatments are assigned at random within each block. Responses are then compared within each block and combined with the responses of other blocks after accounting for the differences between the blocks. When blocks are chosen wisely, it is easier to determine if one treatment is more effective than another.
- A **matched pairs design** is a common form of randomized block design for comparing two treatments. In some matched pairs designs, each subject receives both treatments in a random order. In others, two very similar subjects are paired, and the two treatments are randomly assigned within each pair.

Section 4.2 Exercises

43. Good for the gut? Is fish good for the gut? Researchers tracked 22,000 male physicians for 22 years. Those who reported eating seafood of any kind at least 5 times per week had a 40% lower risk of colon cancer than those who said they ate seafood less than once a week. Explain how confounding makes it unreasonable to conclude that eating seafood causes a reduction in the risk of colon cancer, based on this study.³⁴

44. Straight A's now, healthy later A study by Pamela Herd of the University of Wisconsin–Madison found a link between high school grades and health. Analyzing data from the Wisconsin Longitudinal Study, which has tracked the lives of thousands of Wisconsin high school graduates from the class of 1957, Herd found that students with higher grade-point averages were more likely to say they were in excellent or very good health in their early 60s. Explain how confounding makes it unreasonable to conclude that people will live healthier lives if they increase their GPA, based on this study.³⁵

45. Snacking and TV Does the type of program people watch influence how much they eat? A total of 94 college students were randomly assigned to one of three treatments: watching 20 minutes of a Hollywood action movie (*The Island*), watching the same 20-minute excerpt of the movie with no sound, and watching 20 minutes of an interview program (*Charlie Rose*). While watching, participants were given snacks (M&M'S®, cookies, carrots, and grapes) and allowed to eat as much as they wanted. Subjects who watched the highly stimulating excerpt from *The Island* ate 65% more calories than subjects who watched *Charlie Rose*. Participants who watched the silent version of *The*

Island ate 46% more calories than those who watched *Charlie Rose*.³⁶ Identify the explanatory and response variables in this study. Then determine if it is an experiment or an observational study. Explain your reasoning.

46. 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 student in 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. Identify the explanatory and response variables in this study. Then determine if it is an experiment or an observational study. Explain your reasoning.

47. 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 4½, the more adults tended to rate them, both at age 4½ and at kindergarten, as less likely to get along with others, as more assertive, as disobedient, and as aggressive.”³⁷

- What are the explanatory and response variables?
- Is this an observational study or an experiment? Justify your answer.
- Does this study show that child care makes children more aggressive? Explain your reasoning.

48. 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 using the traits of 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.³⁸

- (a) What are the explanatory and response variables?
- (b) Was this an observational study or an experiment? Justify your answer.
- (c) Does this study show that eating chocolate regularly during pregnancy helps produce infants with good temperament? Explain your reasoning.

49. Growing in the shade The ability to grow in shade may help pine trees 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. Identify the experimental units and the treatments.

50. Sealing your teeth Many children have their molars sealed to help prevent cavities. In an experiment, 120 children aged 6–8 were randomly assigned to a control group, a group in which sealant was applied and reapplied periodically for 36 months, and a group in which fluoride varnish was applied and reapplied periodically for 42 months. After 9 years, the percent of initially healthy molars with cavities was calculated for each group.³⁹ Identify the experimental units and the treatments.

51. 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.

- (a) List the factors in this experiment and state how many levels each factor has.
- (b) If the researchers used every possible combination to form the treatments, how many treatments were included in the experiment?
- (c) List two of the treatments.

52. 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°C or 175°C. Three specimens of fabric will be subjected to each treatment and scored for quality.

- (a) List the factors in this experiment and state how many levels each factor has.
- (b) If the researchers used every possible combination to form the treatments, how many treatments were included in the experiment?
- (c) List two of the treatments.

53. Want a snack? Can snacking on fruit rather than candy reduce later food consumption? Researchers randomly assigned 12 women to eat either 65 calories of berries or 65 calories of candy. Two hours later, all 12 women were given an unlimited amount of pasta to eat. The researchers recorded the amount of pasta consumed by each subject. The women who ate the berries consumed 133 fewer calories, on average. Identify the explanatory and response variables, the experimental units, and the treatments.

54. Pricey pizza? The cost of a meal might affect how customers evaluate and appreciate food. To investigate, researchers worked with an Italian all-you-can-eat buffet to perform an experiment. A total of 139 subjects were randomly assigned to pay either \$4 or \$8 for the buffet and then asked to rate the quality of the pizza on a 9-point scale. Subjects who paid \$8 rated the pizza 11% higher than those who paid only \$4.⁴⁰ Identify the explanatory and response variables, the experimental units, and the treatments.


55. Oils and inflammation The extracts of avocado and soybean oils have been shown to slow cell inflammation in test tubes. Will taking avocado and soybean unsaponifiables (called ASU) help relieve pain for subjects with joint stiffness due to arthritis? In an experiment, 345 men and women were randomly assigned to receive either 300 milligrams of ASU daily for three years or a placebo daily for three years.⁴¹ Explain why it was necessary to include a control group in this experiment.

56. Supplements for testosterone As men age, their testosterone levels gradually decrease. This may cause a reduction in energy, 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.⁴² Explain why it was necessary to include a control group in this experiment.

- 57. Cocoa and blood flow** A study of blood flow involved 27 healthy people aged 18 to 72. Each subject consumed a cocoa beverage containing 900 milligrams of flavonols daily for 5 days. Using a finger cuff, blood flow was measured on the first and fifth days of the study. After 5 days, researchers measured what they called “significant improvement” in blood flow and the function of the cells that line the blood vessels.⁴³ What flaw in the design of this experiment makes it impossible to say if the cocoa really caused the improved blood flow? Explain your answer.

- 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 if the bonus really caused the increase? Explain your answer.


- 59. More oil and inflammation** Refer to Exercise 55.  Could blinding be used in this experiment? Explain your reasoning. Why is blinding an important consideration in this context?

- 60. More testosterone** Refer to Exercise 56. Could blinding be used in this experiment? Explain your reasoning. Why is blinding an important consideration in this context?

- 61. 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 affect the reported results.

- 62. Side effects** Even if an experiment is double-blind, the blinding might be compromised if side effects of the treatments differ. For example, suppose researchers at a skin-care company are comparing their new acne treatment against that of the leading competitor. Fifty subjects are assigned at random to each treatment, and the company’s researchers will rate the improvement for each of the 100 subjects. The researchers aren’t

told which subjects received which treatments, but they know that their new acne treatment causes a slight reddening of the skin. How might this knowledge compromise the blinding? Explain why this is an important consideration in the experiment.

- 63. 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 their performance before the break. Overall, subjects worked harder when told the other student’s dismissal was random.⁴⁴ Describe how you would randomly assign the subjects to the treatments

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

- 64. Precise offers** People often use round prices as first offers in a negotiation. But would a more precise number suggest that the offer was more reasoned and informed? In an experiment, 238 adults played the role of a person selling a used car. Each adult received one of three initial offers: \$2000, \$1865 (a precise under-offer), and \$2135 (a precise over-offer). After hearing the initial offer, each subject made a counter-offer. The difference in the initial offer and counter-offer was the largest in the group that received the \$2000 offer.⁴⁵ Describe how the researchers could have randomly assigned the subjects to the treatments

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

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

66. **Killing weeds** A biologist would like to determine which of two brands of weed killer, X or Y, 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. Then flip a coin. If it lands heads, apply Brand X weed killer to these plants and Brand Y weed killer to the remaining 12 plants. If it lands tails, do the opposite.

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.

67. **Boosting preemies** Do blood-building drugs help brain development in babies born prematurely? Researchers randomly assigned 53 babies, born more than a month premature and weighing less than 3 pounds, to one of three groups. Babies either received injections of erythropoietin (EPO) three times a week, darbepoetin once a week for several weeks, or no treatment. Results? Babies who got the medicines scored much better by age 4 on measures of intelligence, language, and memory than the babies who received no treatment.⁴⁶

- Explain how this experiment used comparison.
- Explain the purpose of randomly assigning the babies to the three treatments.
- Name two variables that were controlled in this experiment and why it was beneficial to control these variables.
- Explain how this experiment used replication. What is the purpose of replication in this context?

68. **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 for over 40 years. 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. Results? Children who were assigned to the preschool program had higher IQ's, higher standardized test scores, and were less likely to repeat a grade in school.⁴⁷

- Explain how this experiment used comparison.
- Explain the purpose of randomly assigning the infants to the two treatments.
- Name two variables that were controlled in this experiment and why it was beneficial to control these variables.
- Explain how this experiment used replication. What is the purpose of replication in this context?

69. **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 what method to use. The study found that patients treated by the new method were more likely to die within 8 years.⁴⁸

- Further study of the data showed that this conclusion was wrong. The extra deaths among patients who were treated with the new method could be explained by other variables. What other variables might be confounded with a doctor's choice of surgical or nonsurgical treatment?
- 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 a completely randomized design for this experiment.

70. **Diet soda and pregnancy** A large study of 3000 Canadian children and their mothers found that the children of mothers who drank diet soda daily during pregnancy were twice as likely to be overweight at age 1 than children of mothers who avoided diet soda during pregnancy.⁴⁹

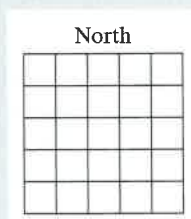
- A newspaper article about this study had the headline "Diet soda, pregnancy: Mix may fuel childhood obesity." This headline suggests that there is a cause-and-effect relationship between diet soda consumption during pregnancy and the weight of the children 1 year after birth. However, this relationship could be explained by other variables. What other variables might be confounded with a mother's consumption of diet soda during pregnancy?
- You have 300 pregnant mothers who are willing to serve as subjects in an experiment that compares three treatments during pregnancy: no diet soda, one diet soda per day, and two diet sodas per day. Write a few sentences describing a completely randomized design for this experiment.

71. **A fruitful experiment** A citrus farmer wants to know which of three fertilizers (A, B, and C) is most effective for increasing the number of oranges on his trees. He

is willing to use 30 mature trees of various sizes from his orchard in an experiment with a randomized block design.

- (a) Describe a randomized block design for this experiment. Justify your choice of blocks.
- (b) Explain why a randomized block design might be preferable to a completely randomized design for this experiment.

- 72. 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) Describe a randomized block design for this experiment. Justify your choice of blocks.
 - (b) Explain why a randomized block design might be preferable to a completely randomized design for this experiment.
- 73. 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.⁵⁰
- (a) Which are the blocks in this experiment: the different diagnoses (asthma, diabetes, or high blood pressure) or the type of care (nurse or doctor)? Why?
 - (b) Explain why a randomized block design is preferable to a completely randomized design in this context.
 - (c) Suppose the experiment used only diabetes patients, but there were still 1316 subjects willing to participate. What advantage would this offer? What disadvantage?
- 74. 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 three 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?

- 75. 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, then genetics and type of diet would be confounded. Explain this statement carefully.
- (b) Describe how to randomly assign the rats to treatments using a randomized block design with litters as blocks.
- (c) Use technology or Table D to carry out the random assignment.

- 76. Comparing weight-loss treatments** A total of 20 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.

- (a) If the 5 most overweight women were assigned Treatment A, the next 5 most overweight women were assigned Treatment B, and so on, then the amount overweight and type of treatment would be confounded. Explain this statement carefully.
- (b) Describe how to randomly assign the women to treatments using a randomized block design. Use blocks of size 4 formed by the amount overweight.
- (c) Use technology or Table D to carry out the random assignment.

77. SAT preparation A school counselor wants to compare the effectiveness of an online SAT preparation program with an in-person SAT preparation class. For an experiment, the counselor recruits 30 students who have already taken the SAT once. The response variable will be the improvement in SAT score.

- (a) Design an experiment that uses a completely randomized design to investigate this question.
- (b) Design an experiment that uses a matched pairs design to investigate this question. Explain your method of pairing.
- (c) Which design do you prefer? Explain your answer.

78. Valve surgery Medical researchers want to compare the success rate of a new non-invasive method for replacing heart valves using a cardiac catheter with traditional open-heart surgery. They have 40 male patients, ranging in age from 55 to 75, who need valve replacement. One of several response variables will be the percentage of blood that flows backward—in the wrong direction—through the valve on each heartbeat.

- (a) Design an experiment that uses a completely randomized design to investigate this question.
- (b) Design an experiment that uses a matched pairs design to investigate this question. Explain your method of pairing.
- (c) Which design do you prefer? Explain your answer.

79. 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. Each subject will complete two sessions in the simulator: one while talking on the hands-free phone and the other while just driving. The order of the two sessions for each subject will be determined at random. The route, driving conditions, and traffic flow will be the same in both sessions.

- (a) What type of design did the researchers use in their study?
- (b) Explain why the researchers chose this design instead of a completely randomized design.
- (c) Why is it important to randomly assign the order of the treatments?
- (d) Explain how and why researchers controlled for other variables in this experiment.

80. Chocolate gets my heart pumping Cardiologists at Athens Medical School in Greece wanted to test if chocolate affects blood vessel function. 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 received the other treatment. The order in which subjects received the bittersweet and fake chocolate was determined at random. The subjects had no chocolate outside the study, and investigators didn't know if a subject had eaten the real or the fake chocolate. An ultrasound was taken of each volunteer's upper arm to observe the functioning of the cells in the walls of the main artery. The researchers 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).⁵¹

- (a) What type of design did the researchers use in their study?
- (b) Explain why the researchers 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?
- (d) Explain how and why researchers controlled for other variables in this experiment.

81. 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, because male and female bodies might respond differently to deodorant. Give all the males Brand A deodorant and all the females Brand B. Have the principal judge 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 part (a).

82. Close shave Which of two brands (X or Y) of electric razor shaves closer? Researchers want to design and carry out an experiment to answer this question using 50 adult male volunteers. Here's one idea: Have all 50 subjects shave the left sides of their faces with the Brand X razor and shave the right sides of their faces with the Brand Y razor. Then have each man decide which razor gave the closer shave and compile the results.

- (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 part (a).

Multiple Choice Select the best answer for Exercises 83–90.

83. Can a vegetarian or low-salt diet reduce blood pressure? Men with high blood pressure are assigned at random to one of 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

- (a) one factor, the type of diet.
- (b) two factors, high blood pressure and type of diet.
- (c) two factors, normal/vegetarian diet and unrestricted/restricted salt.
- (d) three factors, men, high blood pressure, and type of diet.
- (e) four factors, the four diets being compared.

84. In the experiment of the preceding exercise, the subjects were randomly assigned to the different treatments. What is the most important reason for this random assignment?

- (a) Random assignment eliminates the effects of other variables such as stress and body weight.
- (b) Random assignment is a good way to create groups of subjects that are roughly equivalent at the beginning of the experiment.
- (c) Random assignment makes it possible to make a conclusion about all men.
- (d) Random assignment reduces the amount of variation in blood pressure.
- (e) Random assignment prevents the placebo effect from ruining the results of the study.

85. To investigate if standing up while studying affects performance in an algebra class, a teacher assigns half of the 30 students in his class to stand up while studying and assigns the other half to not stand up while studying. To determine who receives which treatment, the teacher identifies the two students who did best on the last exam and randomly assigns one to stand and one to not stand. The teacher does the same for the next two highest-scoring students and continues in this manner until each student is assigned a treatment. Which of the following best describes this plan?

- (a) This is an observational study.
- (b) This is an experiment with blocking.
- (c) This is a completely randomized experiment.
- (d) This is a stratified random sample.
- (e) This is a cluster sample.

86. A gardener wants to try different combinations of fertilizer (none, 1 cup, 2 cups) and mulch (none,

wood chips, pine needles, plastic) to determine which combination produces the highest yield for a variety of green beans. He has 60 green-bean plants to use in the experiment. If he wants an equal number of plants to be assigned to each treatment, how many plants will be assigned to each treatment?

- (a) 1 (b) 3 (c) 4
- (d) 5 (e) 12

87. 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. Based on these results, is it reasonable to conclude that corn variety 1 is more productive than corn variety 2?

- (a) Yes, because 140 bushels per acre is greater than 110 bushels per acre.
- (b) Yes, because the study was done at a research farm.
- (c) No, because there may be other differences between the two years besides the corn variety.
- (d) No, because there was no use of a placebo in the experiment.
- (e) No, because the experiment wasn't double-blind.

88. A report in a medical journal notes that the risk of developing Alzheimer's disease among subjects who regularly opted to take the drug ibuprofen was about half the risk of those who did not. Is this good evidence that ibuprofen is effective in preventing Alzheimer's disease?

- (a) Yes, because the study was a randomized, comparative experiment.
- (b) No, because the effect of ibuprofen is confounded with the placebo effect.
- (c) Yes, because the results were published in a reputable professional journal.
- (d) No, because this is an observational study. An experiment would be needed to confirm (or not confirm) the observed effect.
- (e) Yes, because a 50% reduction can't happen just by chance.

89. A farmer is conducting an experiment to determine which variety of apple tree, Fuji or Gala, will produce more fruit in his orchard. The orchard is divided into 20 equally sized square plots. He has 10 trees of each variety and randomly assigns each tree to a separate plot in the orchard. What are the experimental unit(s) in this study?

- (a) The trees (b) The plots (c) The apples
- (d) The farmer (e) The orchard

90. Two essential features of all statistically designed experiments are
- comparing several treatments; using the double-blind method.
 - comparing several treatments; using chance to assign subjects to treatments.
 - always having a placebo group; using the double-blind method.
 - using a block design; using chance to assign subjects to treatments.
 - using enough subjects; always having a control group.

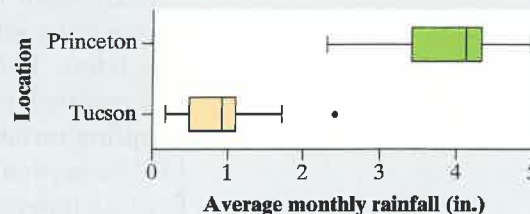
Recycle and Review

91. **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.

- What percent of seeds weigh more than 500 mg?
- If we discard the lightest 10% of these seeds, what is the smallest weight among the remaining seeds?

92. **Comparing rainfall (1.3)** The boxplots summarize the distributions of average monthly rainfall (in inches) for Tucson, Arizona, and Princeton, New Jersey.⁵² Compare these distributions.



SECTION 4.3 Using Studies Wisely

LEARNING TARGETS *By the end of the section, you should be able to:*

- Explain the concept of sampling variability when making an inference about a population and how sample size affects sampling variability.
- Explain the meaning of statistically significant in the context of an experiment and use simulation to determine if the results of an experiment are statistically significant.
- Identify when it is appropriate to make an inference about a population and when it is appropriate to make an inference about cause and effect.
- Evaluate if a statistical study has been carried out in an ethical manner.*

Researchers who conduct statistical studies often want to draw conclusions 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 scientists hoped to show that sleep deprivation causes a decrease in performance two days later.⁵³

What conclusions can be drawn from a particular study? The answer depends on how the data were collected.

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

Inference for Sampling

When the members of a sample are selected at random from a population, we can use the sample results to *infer* things about the population. That is, we can make *inferences* about the population from which the sample was randomly selected. Inference from convenience samples or voluntary response samples would be misleading because these methods of choosing a sample are biased. In these cases, we are almost certain that the sample does *not* fairly represent the population.

Even when making an inference from a random sample, it would be surprising if the estimate from the sample was exactly equal to the truth about the population. For example, in a random sample of 1399 U.S. teens aged 13–18, 26% reported more than 8 hours of entertainment media use per day. Because of **sampling variability**, it would be surprising if exactly 26% of *all* U.S. teens aged 13–18 reported more than 8 hours of entertainment media use per day. Why? Because different samples of 1399 U.S. teens aged 13–18 will include different sets of people and produce different estimates.

DEFINITION Sampling variability

Sampling variability refers to the fact that different random samples of the same size from the same population produce different estimates.

The following activity explores the idea of sampling variability.

ACTIVITY

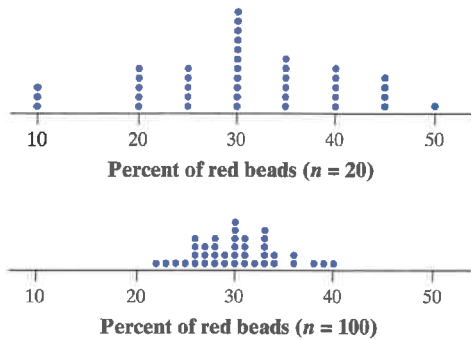
Exploring sampling variability

When making an inference about a population from a random sample, we shouldn't expect the estimate to be exactly correct. But how much do sample results vary? Your teacher has prepared a large population of beads, where 30% have a certain color (e.g., red) so you can explore this question.



G. Curt Fiedler/Getty Images

1. In a moment, you will select a random sample of 20 beads. Do you expect that your sample will contain exactly 30% red beads? Explain your reasoning.
2. Mix the beads thoroughly, select a random sample of 20 beads from the population, calculate the percent of red beads in the sample, and replace the beads in the population.
3. After all students have selected a sample, make a class dotplot showing each student's estimate for the percent of red beads. Where is the graph centered? How much does the percent of red beads vary?
4. Imagine that you repeated Steps 2 and 3 with random samples of size 100. How do you expect the dotplots would compare? *If there's time, select random samples of size 100 to confirm your answer.*



When Mrs. Storrs's class of 40 students did the red bead activity, they produced the dotplot shown at left (top) for samples of size 20. The dotplot is centered around 30%, the true percent of red beads. This shouldn't be surprising because random sampling helps avoid bias. Notice also that the estimates varied from 10% to 50% and that only 11 of the 40 estimates were equal to exactly 30%.

To see the effect of increasing the sample size, we simulated 40 random samples of size 100 from the same population and recorded the percent of red beads in each sample. Notice that the graph is still centered at 30%, but there is much less variability.

SAMPLING VARIABILITY AND SAMPLE SIZE

Larger random samples tend to produce estimates that are closer to the true population value than smaller random samples. In other words, estimates from larger samples are more precise.

EXAMPLE

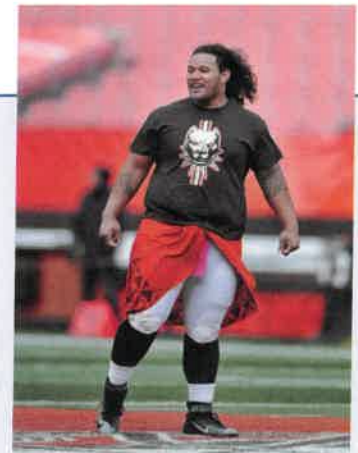
Weighing football players Inference for sampling

PROBLEM: How much do National Football League (NFL) players weigh, on average? In a random sample of 50 NFL players, the average weight is 244.4 pounds.

- Do you think that 244.4 pounds is the true average weight of all NFL players? Explain your answer.
- Which would be more likely to give an estimate close to the true average weight of all NFL players: a random sample of 50 players or a random sample of 100 players? Explain your answer.

SOLUTION:

- No. Different samples of size 50 would produce different average weights. So it would be surprising if this estimate is equal to the true average weight of all NFL players.
- A random sample of 100 players, because estimates tend to be closer to the truth when the sample size is larger.



Diamond Images/Getty Images

FOR PRACTICE, TRY EXERCISE 93

Estimates from random samples often come with a *margin of error* that allows us to create an interval of plausible values for the true population value. In the preceding example about NFL players, the margin of error for the estimate of 244.4 pounds is 14.2 pounds. Based on this margin of error, it wouldn't be surprising if the true average weight for all NFL players was as small as $244.4 - 14.2 = 230.2$ pounds or as large as $244.4 + 14.2 = 258.6$ pounds.

You will learn how to calculate the margin of error in Chapter 8. For now, make sure to remember the effect of sampling variability when using data from a random sample to make an inference about a population.

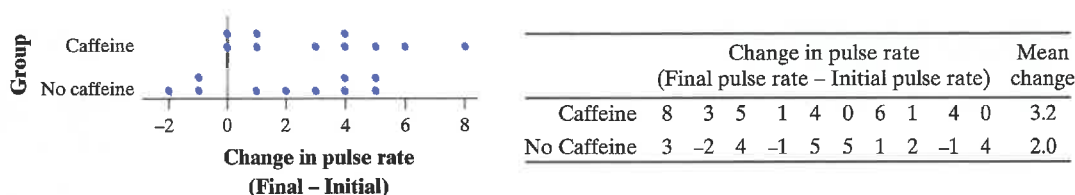
Inference for Experiments

Well-designed experiments allow for inferences about cause and effect. But we should only conclude that changes in the explanatory variable cause changes in the response variable if the results of an experiment are **statistically significant**.

DEFINITION Statistically significant

When the observed results of a study are too unusual to be explained by chance alone, the results are called **statistically significant**.

Mr. Wilcox and his students decided to perform the caffeine experiment from the preceding section. In their experiment, 10 student volunteers were randomly assigned to drink cola with caffeine and the remaining 10 students were assigned to drink caffeine-free cola. The table and graph show the change in pulse rate for each student (Final pulse rate – Initial pulse rate), along with the mean change for each group.



The dotplots provide some evidence that caffeine has an effect on pulse rates. The mean change for the 10 students who drank cola with caffeine was 3.2, which is 1.2 greater than for the group who drank caffeine-free cola. But are the results statistically significant?

Recall that the purpose of random assignment in this experiment was to create two groups that were roughly equivalent at the beginning of the experiment. Subjects with high caffeine tolerance should be split up in about equal numbers, subjects with high metabolism should be split up in about equal numbers, and so on.

Of course, the random assignment is unlikely to produce groups that are exactly equivalent. One group might get more “favorable” subjects just by chance. That is, the caffeine group might end up with a few extra subjects who were likely to have a pulse rate increase, just due to chance variation in the random assignment.

There are two ways to explain why the mean change in pulse rate was 1.2 greater for the caffeine group:

1. Caffeine does *not* have an effect on pulse rates, and the difference of 1.2 happened because of chance variation in the random assignment.
2. Caffeine increases pulse rates.

If it is plausible to get a difference of 1.2 or more simply due to the chance variation in random assignment, the results of the experiment are not statistically

significant. But if it is very unlikely to get a difference of 1.2 or more by chance alone, we rule out Explanation 1 and say the results are statistically significant—and that caffeine increases pulse rates.

How can we determine if a difference of 1.2 is statistically significant? You'll find out in the following activity.

ACTIVITY

Analyzing the caffeine experiment



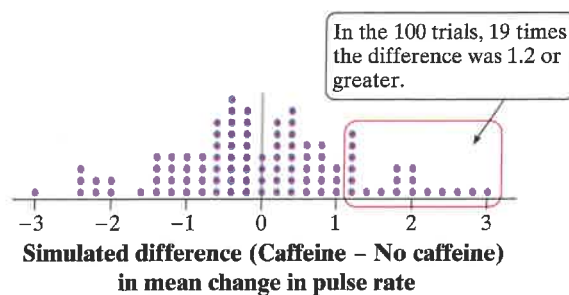
In the experiment performed by Mr. Wilcox's class, the mean change in pulse rate for the caffeine group was 1.2 greater than the mean change for the no-caffeine group. This provides some evidence that caffeine increases pulse rates. Is this evidence convincing? Or is it plausible that a difference of 1.2 would arise just due to chance variation in the random assignment? In this activity, we'll investigate by seeing what differences typically occur just by chance, assuming caffeine doesn't affect pulse rates. That is, we'll assume that the change in pulse rate for a particular student would be the same regardless of what treatment he or she was assigned.

	Change in pulse rate (Final pulse rate – Initial pulse rate)										Mean change
Caffeine	8	3	5	1	4	0	6	1	4	0	3.2
No Caffeine	3	-2	4	-1	5	5	1	2	-1	4	2.0

1. Gather 20 index cards to represent the 20 students in this experiment. On each card, write one of the 20 outcomes listed in the table. For example, write "8" on the first card, "3" on the second card, and so on.
2. Shuffle the cards and deal two piles of 10 cards each. This represents randomly assigning the 20 students to the two treatments, *assuming that the treatment received doesn't affect the change in pulse rate*. The first pile of 10 cards represents the caffeine group, and the second pile of 10 cards represents the no-caffeine group.
3. Find the mean change for each group and subtract the means (Caffeine – No caffeine). *Note:* It is possible to get a negative difference.
4. Your teacher will draw and label an axis for a class dotplot. Plot the difference you got in Step 3 on the graph.
5. In Mr. Wilcox's class, the observed difference in means was 1.2. Is a difference of 1.2 statistically significant? Discuss with your classmates.

We used technology to perform 100 trials of the simulation described in the activity. The dotplot in Figure 4.3 shows that getting a difference of 1.2 isn't that unusual. In 19 of the 100 trials, we obtained a difference of 1.2 or more simply due to chance variation in the random assignment.

FIGURE 4.3 Dotplot showing the differences in means that occurred in 100 simulated random assignments, assuming that caffeine has no effect on pulse rates.

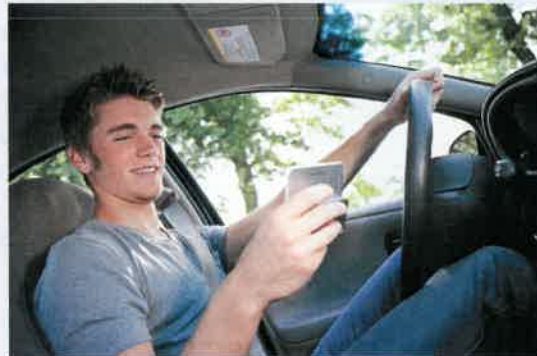


Because the difference of 1.2 or greater is somewhat likely to occur by chance alone, the results of Mr. Wilcox's class experiment aren't statistically significant. Based on this experiment, there isn't convincing evidence that caffeine increases pulse rates.

EXAMPLE

Distracted driving Inference for experiments

PROBLEM: 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 or not the driver stopped at a rest area that was specified by researchers before the simulation started. The table shows the results.⁵⁴



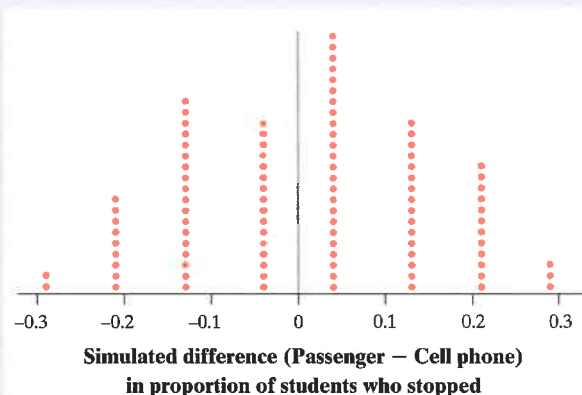
Sean Locke Photography/Shutterstock.com

- (a) Calculate the difference (Passenger – Cell phone) in the proportion of students who stopped at the rest area in the two groups.

One hundred trials of a simulation were performed to see what differences in proportions would occur due only to chance variation in the random assignment, assuming that the type of distraction did not affect whether a subject stopped at the rest area. That is, 33 “stoppers” and 15 “non-stoppers” were randomly assigned to two groups of 24.

- (b) There are three dots at 0.29. Explain what these dots mean in this context.
- (c) Use the results of the simulation to determine if the difference in proportions from part (a) is statistically significant. Explain your reasoning.

		Treatment		
		Cell phone	Passenger	Total
Response	Stopped at rest area	12	21	33
	Didn't stop	12	3	15
	Total	24	24	48



SOLUTION:

- (a) $\text{Difference in proportions} = 21/24 - 12/24 = 0.875 - 0.500 = 0.375$
- (b) When we assumed that the type of distraction doesn't matter, there were three simulated random assignments where the difference in the proportion of students who stopped at the rest area was 0.29.
- (c) Because a difference of 0.375 or greater never occurred in the simulation, the difference is statistically significant. It is extremely unlikely to get a difference this big simply due to chance variation in the random assignment.

Because the difference is statistically significant, we can make a cause-and-effect conclusion: talking on a cell phone is more distracting than talking with a passenger—at least for subjects like those in the experiment.

In the caffeine example, we said that a difference in means of 1.2 was not unusual because a difference that big or bigger occurred 19% of the time by chance alone. In the distracted drivers example, we said that a difference in proportions of 0.375 was unusual because a difference this big or bigger occurred 0% of the time by chance alone. So the boundary between “not unusual” and “unusual” must be somewhere between 0% and 19%. For now, we recommend using a boundary of 5% so that differences that would occur less than 5% of the time by chance alone are considered statistically significant.

The Scope of Inference: Putting it All Together

The type of conclusion that can be drawn from a study depends on how the data in the study were collected.

In the example about average weight in the NFL, the players were *randomly selected* from all NFL players. As you learned in Section 4.1, random sampling helps to avoid bias and produces reliable estimates of the truth about the population. Because the mean weight in the sample of players was 244.4 pounds, our best guess for the mean weight in the population of all NFL players is 244.4 pounds. Even though our estimates are rarely exactly correct, when samples are selected at random, we can make an *inference about the population*.

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.

In the distracted driver experiment, subjects were *randomly assigned* to talk on a cell phone or talk to a passenger. As you learned in Section 4.2, random assignment helps ensure that the two groups of subjects are as alike as possible before the treatments are imposed. If the group assigned to talk with a passenger remembers to stop at the rest area more often than the group assigned to talk on a cell phone, 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 researchers could safely conclude that talking on a cell phone is more distracting than talking to a passenger. That is, they can make an *inference about cause and effect*. However, because the experiment used volunteer subjects, the scientists can only apply this conclusion to subjects like the ones in their experiment.

Let's recap what we've learned about the scope of inference in a statistical study.

THE SCOPE OF INFERENCE

- Random selection of individuals allows inference about the population from which the individuals were chosen.
- Random assignment of individuals to groups allows inference about cause and effect.

The following chart summarizes the possibilities.⁵⁵

		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 for individuals like those who received the treatments. Observational studies don't randomly assign individuals to groups, which makes it challenging to make an inference about cause and effect. But an observational study that uses random sampling can make an inference about the population.

EXAMPLE

When will I ever use this stuff?
The scope of inference

PROBLEM: 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 central idea of the program is that teachers show students how the topics they're learning about can be applied to 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, determined at random, used CareerStart along with the district's standard curriculum. The other 7 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 or not the students graduated from high school.

Results: Students at schools that used CareerStart generally had significantly better attendance and fewer discipline problems, earned higher test scores, reported greater engagement in their classes, and were more likely to graduate.⁵⁶

What conclusion can we draw from this study? Explain your reasoning.

SOLUTION:

Because treatments were randomly assigned and the results were significant, we can conclude that using the CareerStart curriculum caused better attendance, fewer discipline problems, higher test scores, greater engagement, and increased graduation rates. However, these results only apply to schools like those in the study because the schools were not randomly selected from any population.

With no random selection, the results of the study should be applied only to schools like those in the study. With random assignment, it is possible to make an inference about cause and effect.

FOR PRACTICE, TRY EXERCISE 103



CHECK YOUR UNDERSTANDING

When an athlete suffers a sports-related concussion, does it help to remove the athlete from play immediately? Researchers recruited 95 athletes seeking care for a sports-related concussion at a medical clinic and followed their progress during recovery. Researchers found statistically significant evidence that athletes who were removed from play immediately recovered more quickly, on average, than athletes who continued to play.⁵⁷ What conclusion can we draw from this study? Explain your answer.

The Challenges of Establishing Causation

A well-designed experiment can tell 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. In some cases, it isn't practical or even ethical to do an experiment. Consider these important questions:

- 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 attend church or to smoke cigarettes. The best data we have about these and many other cause-and-effect questions come from 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 some other variable? Is there some genetic factor that makes people both more likely to become 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. Still, 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.

There are several 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 some other 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.

archives/Getty Images



- *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 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.”⁵⁸ 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*

There are some potential experiments that are clearly unethical. In other cases, the boundary between “ethical” and “unethical” isn’t as clear. Decide if 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.
- 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. The law requires that studies carried out or funded by the federal government obey these principles.⁵⁹ But neither the law nor the consensus of experts is completely clear about the details of their application.

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

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.

Institutional Review Boards The purpose of an *institutional review board* is not to decide whether a proposed study will produce valuable information or if 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.

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 questionnaire, 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 roughly 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.

Confidentiality 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 believed 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.

Section 4.3**Summary**

- **Sampling variability** refers to the idea that different random samples of the same size from the same population produce different estimates. Reduce sampling variability by increasing the sample size.
- When the observed results of a study are too unusual to be explained by chance alone, we say that the results are **statistically significant**.
- Most studies aim to make inferences that go beyond the data produced.

- **Inference about a population** requires that the individuals taking part in a study be randomly selected from the population.
- A well-designed experiment that randomly assigns experimental units to treatments allows **inference about cause and effect**.
- 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 other 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**.

Section 4.3 Exercises

93. Tweet, tweet! What proportion of students at your school use Twitter? To find out, you survey a simple random sample of students from the school roster.

- (a) Will your sample result be exactly the same as the true population proportion? Explain your answer.
- (b) Which would be more likely to produce a sample result closer to the true population value: an SRS of 50 students or an SRS of 100 students? Explain your answer.

94. Far from home? A researcher wants to estimate the average distance that students at a large community college live from campus. To find out, she surveys a simple random sample of students from the registrar's database.

- (a) Will the researcher's sample result be exactly the same as the true population mean? Explain your answer.
- (b) Which would be more likely to produce a sample result closer to the true population value: an SRS of 100 students or an SRS of 50 students? Explain your answer.

95. Football on TV A Gallup poll conducted telephone interviews with a random sample of 1000 adults aged 18 and older. Of these, 37% said that football is their favorite sport to watch on television. The margin of error for this estimate is 3.1 percentage points.

- (a) Would you be surprised if a census revealed that 50% of adults in the population would say their favorite sport to watch on TV is football? Explain your answer.
- (b) Explain how Gallup could decrease the margin of error.

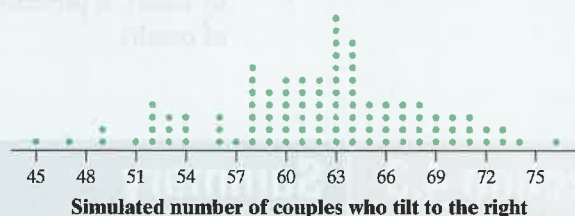
96. Car colors in Miami Using a webcam, a traffic analyst selected a random sample of 800 cars traveling on I-95 in Miami on a weekday morning. Among the 800 cars

in the sample, 24% were white. The margin of error for this estimate is 3.0 percentage points.

- (a) Would you be surprised if a census revealed that 26% of cars on I-95 in Miami on a weekday morning were white? Explain your answer.
- (b) Explain how the traffic analyst could decrease the margin of error.

97. Kissing the right way According to a newspaper article, "Most people are kissing the 'right way.'" That is, according to a study, the majority of couples prefer to tilt their heads to the right when kissing. In the study, a researcher observed a random sample of 124 kissing couples and found that 83/124 (66.9%) of the couples tilted to the right.⁶⁰ To determine if these data provide convincing evidence that couples are more likely to tilt their heads to the right, 100 simulated SRSs were selected.

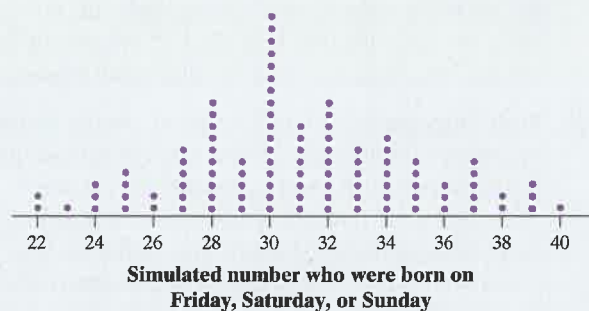
Each dot in the graph shows the number of couples that tilt to the right in a simulated SRS of 124 couples, assuming that each couple has a 50% chance of tilting to the right.



- (a) Explain how the graph illustrates the concept of sampling variability.
- (b) Based on the data from the study and the results of the simulation, is there convincing evidence that couples prefer to kiss the "right way"? Explain your answer.

98. **Weekend birthdays** Over the years, the percentage of births that are planned caesarean sections has been rising. Because doctors can schedule these deliveries, there might be more children born during the week and fewer born on the weekend than if births were uniformly distributed throughout the week. To investigate, Mrs. McDonald and her class selected an SRS of 73 people born since 1993. Of these people, 24 were born on Friday, Saturday, or Sunday.

To determine if these data provide convincing evidence that fewer than 43% (3/7) of people born since 1993 were born on Friday, Saturday, or Sunday, 100 simulated SRSs were selected. Each dot in the graph shows the number of people that were born on Friday, Saturday, or Sunday in a simulated SRS of 73 people, assuming that each person had a 43% chance of being born on one of these three days.

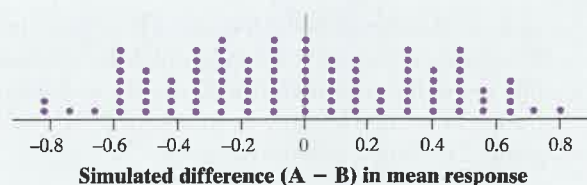


- (a) Explain how the graph illustrates the concept of sampling variability.
- (b) Based on the data from the study and the results of the simulation, is there convincing evidence that fewer than 43% of people born since 1993 were born on Friday, Saturday, or Sunday? Explain your answer.

99. **I work out a lot** Are people influenced by what others say? Michael conducted an experiment in front of a popular gym. As people entered, he asked them how many days they typically work out per week. As he asked the question, he showed the subjects one of two clipboards, determined at random. Clipboard A had the question and many responses written down, where the majority of responses were 6 or 7 days per week. Clipboard B was the same, except most of the responses were 1 or 2 days per week. The mean response for the Clipboard A group was 4.68 and the mean response for the Clipboard B group was 4.21.⁶¹

- (a) Calculate the difference (Clipboard A – Clipboard B) in the mean number of days for the two groups.

One hundred trials of a simulation were performed to see what differences in means would occur due only to chance variation in the random assignment, assuming that the responses on the clipboard don't matter. The results are shown in the dotplot.

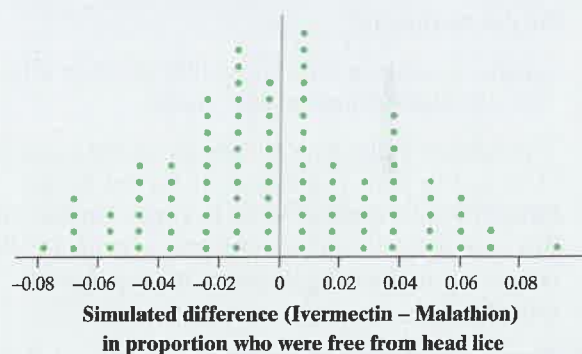


- (b) There is one dot at 0.72. Explain what this dot means in this context.
- (c) Use the results of the simulation to determine if the difference in means from part (a) is statistically significant. Explain your reasoning.

100. **A louse-y situation** A study published in the *New England Journal of Medicine* compared two medicines to treat head lice: an oral medication called ivermectin and a topical lotion containing malathion. Researchers studied 812 people in 376 households in seven areas around the world. Of the 185 households randomly assigned to ivermectin, 171 were free from head lice after 2 weeks, compared with only 151 of the 191 households randomly assigned to malathion.⁶²

- (a) Calculate the difference (Ivermectin – Malathion) in the proportion of households that were free from head lice in the two groups.

One hundred trials of a simulation were performed to see what differences in proportions would occur due only to chance variation in the random assignment, assuming that the type of medication doesn't matter. The results are shown in the dotplot.



- (b) There is one dot at 0.09. Explain what this dot means in this context.
- (c) Use the results of the simulation to determine if the difference in proportions from part (a) is statistically significant. Explain your reasoning.

101. **Acupuncture and pregnancy** A study sought to determine if the ancient Chinese art of acupuncture could help infertile women become pregnant.⁶³ A total of 160 healthy women undergoing assisted reproductive therapy 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.

- (a) Why did researchers randomly assign the subjects to the two treatments?
- (b) 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.
- (c) Explain why the design of the study prevents us from concluding that acupuncture caused the difference in pregnancy rates.

102. Do 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.⁶⁴

- (a) Why did researchers randomly assign the subjects to the diet treatments?
- (b) Explain to someone who knows little statistics what "lost significantly more weight" means.
- (c) 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 in a year. Explain how this information could be consistent with the fact that weight loss in the two groups was not significantly different.

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 statistically significant evidence that they do. 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.⁶⁵ (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 your reasoning.

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 significantly higher average charge. What conclusion can we draw from this study? Explain your reasoning.

105. 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 significantly more likely to have died."⁶⁶ What conclusion can we draw from this study? Explain your reasoning.

106. Rude surgeons Is a friendly surgeon a better surgeon? In a study of more than 32,000 surgical patients from 7 different medical centers, researchers classified surgeons by the number of unsolicited complaints that had been recorded about their behavior. The researchers found that surgical complications were significantly more common in patients whose surgeons had received the most complaints, compared with patients whose surgeons had received the fewest complaints.⁶⁷ What conclusion can we draw from this study? Explain your reasoning.

107. Berry good Eating blueberries and strawberries might improve heart health, according to a long-term study of 93,600 women who volunteered to take part. These berries are high in anthocyanins due to their pigment. Women who reported consuming the most anthocyanins had a significantly smaller risk of heart attack compared to the women who reported consuming the least. What conclusion can we draw from this study? Explain your reasoning.⁶⁸

108. Exercise and memory A study of strength training and memory randomly assigned 46 young adults to two groups. After both groups were shown 90 pictures, one group had to bend and extend one leg against heavy resistance 60 times. The other group stayed relaxed, while the researchers used the same exercise machine to bend and extend their legs with no resistance. Two days later, each subject was shown 180 pictures—the original 90 pictures plus 90 new pictures and asked to identify which pictures were shown two days earlier. The resistance group was significantly more successful in identifying these pictures than was the relax group. What conclusions can we draw from this study? Explain your reasoning.⁶⁹

- 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"?
- (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 religious leader? An activist for patients' rights?)
- 111.* Facebook emotions** In cooperation with researchers from Cornell University, Facebook randomly selected almost 700,000 users for an experiment in "emotional contagion." Users' news feeds were manipulated (without their knowledge) to selectively show postings from their friends that were either more positive or more negative in tone, and the emotional tone of their own subsequent postings was measured. The researchers found evidence that people who read emotionally negative postings were more likely to post messages with a negative tone, whereas those who read positive messages were more likely to post messages with a positive tone. The research was widely criticized for being unethical. Explain why.⁷⁰
- 112.* No consent needed?** In which of the circumstances listed here 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 its 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 its members.
- 113.* Anonymous? Confidential?** One of the most important nongovernment surveys in the United States is the National Opinion Research Center's General Social Survey (GSS). 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 anonymous or confidential screening? Why?
- 115.* The Willowbrook hepatitis studies** In the 1960s, children entering the Willowbrook State School, an institution for the intellectually disabled on Staten Island in New York, 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 believed that providing these services to some but not other persons in the same institution raised ethical questions. Do you agree?

⁷⁰Exercises 109–116: This is an important topic, but it is not required for the AP[®] Statistics exam.

Multiple Choice Select the best answer for Exercises 117 and 118.

117. Do product labels influence customer perceptions? To find out, researchers recruited more than 500 adults and asked them to estimate the number of calories, amount of added sugar, and amount of fat in a variety of food products. Half of the subjects were randomly assigned to evaluate products with the word “Natural” on the label, while the other half were assigned to evaluate the same products without the “Natural” label. On average, the products with the “Natural” label were judged to have significantly fewer calories. Based on this study, is it reasonable to conclude that including the word “Natural” on the label causes a reduction in estimated calories?

- (a) No, because the adults weren’t randomly selected from the population of all adults.
- (b) No, because there wasn’t a control group for comparison.
- (c) No, because association doesn’t imply causation.
- (d) Yes, because the adults were randomly assigned to the treatments.
- (e) Yes, because there were a large number of adults involved in the study.

118. Some news organizations maintain a database of customers who have volunteered to share their opinions on a variety of issues. Suppose that one of these databases includes 9000 registered voters in California. To measure the amount of support for a controversial ballot issue, 1000 registered voters in California are randomly selected from the database and asked their opinion. Which of the following is the largest population to which the results of this survey should be generalized?

- (a) The 1000 people in the sample
- (b) The 9000 registered voters from California in the database

- (c) All registered voters in California
- (d) All California residents
- (e) All registered voters in the United States

Review and Recycle

119. 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:

Opinion about using animals for medical testing	Gender			
		Male	Female	Total
	Strongly agree	76	59	135
	Agree	270	247	517
	Neither agree nor disagree	87	139	226
	Disagree	61	123	184
	Strongly agree	22	68	90
	Total	516	636	1152

- (a) Construct segmented bar graphs to display the distribution of opinion for males and for females.
- (b) Is there an association between gender and opinion for the members of this sample? Explain your answer.

120. 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%.⁷¹ Explain how this difference could happen.

Chapter 4 Wrap-Up



FRAPPY! FREE RESPONSE AP[®] PROBLEM, YAY!

The following problem is modeled after actual AP[®] Statistics exam free-response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

In a recent study, 166 adults from the St. Louis area were recruited and randomly assigned to receive one of two treatments for a sinus infection. Half of the subjects received an antibiotic (amoxicillin) and the other half received a placebo.⁷²

- (a) Describe how the researchers could have assigned treatments to subjects if they wanted to use a completely randomized design.
- (b) All the subjects in the experiment had moderate, severe, or very severe symptoms at the beginning of the study. Describe one statistical benefit and one statistical drawback for using subjects with

moderate, severe, or very severe symptoms instead of just using subjects with very severe symptoms.

- (c) At different stages during the next month, all subjects took the sino-nasal outcome test. After 10 days, the difference in average test scores was *not* statistically significant. In this context, explain what it means for the difference to be not statistically significant.
- (d) One possible way that researchers could have improved the study is to use a randomized block design. Explain how the researchers could have incorporated blocking in their design.

After you finish, you can view two example solutions on the book's website (highschool.bfwpub.com/tps6e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score

Chapter 4 Review

Section 4.1: Sampling and Surveys

In this section, you learned that a population is the group of all individuals that we want information about. A sample is the subset of the population that we use to gather this information. The goal of most sample surveys is to use information from the sample to draw conclusions about the population. Choosing people for a sample because they are located nearby or letting people choose whether or not to be in the sample are poor ways to select a sample. Because convenience sampling and voluntary response sampling will produce estimates that are likely to underestimate or likely to overestimate the value you want to know, these methods of choosing a sample are biased.

To avoid bias in the way the sample is formed, the members of the sample should be chosen at random. One way to do this is with a simple random sample (SRS), which is equivalent to selecting well-mixed slips of paper from a hat. It is often more convenient to select an SRS using technology or a table of random digits.

Two other random sampling methods are stratified sampling and cluster sampling. To obtain a stratified random sample, divide the population into groups (strata) of individuals that are likely to have similar responses, take an SRS from each stratum, and combine the chosen individuals to form the sample. Stratified random samples can produce estimates with much greater precision than simple random

samples. To obtain a cluster sample, divide the population into groups (clusters) of individuals that are in similar locations, randomly select clusters, and use every individual in the chosen clusters. Cluster samples are easier to obtain than simple random samples or stratified random samples, but they may not produce very precise estimates.

Finally, you learned about other issues in sample surveys that can lead to bias: undercoverage occurs when the sampling method systematically underrepresents one part of the population. Nonresponse describes when answers cannot be obtained from some people that were chosen to be in the sample. Bias can also result when some people in the sample don't give accurate responses due to question wording, interviewer characteristics, or other factors.

Section 4.2: Experiments

In this section, you learned about the difference between observational studies and experiments. Experiments deliberately impose a treatment to see if there is a cause-and-effect relationship between two variables. Observational studies look at relationships between two variables, but make it difficult to show cause and effect because other variables may be confounded with the explanatory variable. Variables are confounded when it is impossible to determine which of the variables is causing a change in the response variable.

A common type of comparative experiment uses a completely randomized design. In this type of design, the experimental units are assigned to the treatments at random. With random assignment, the treatment groups will be roughly equivalent at the beginning of the experiment. Replication means giving each treatment to as many experimental units as possible. This makes it easier to see the effects of the treatments because the effects of other variables are more likely to be balanced among the treatment groups.

During an experiment, it is important that other variables be controlled (kept the same) for each experimental unit. Doing so helps avoid confounding and removes a possible

source of variation in the response variable. Also, beware of the placebo effect—the tendency for people to improve because they expect to, not because of the treatment they are receiving. One way to make sure that all experimental units have the same expectations is to make them blind—unaware of which treatment they are receiving. When the people interacting with the subjects and measuring the response variable are also blind, the experiment is called double-blind.

Blocking in experiments is similar to stratifying in sampling. To form blocks, group together experimental units that are similar with respect to a variable that is associated with the response. Then randomly assign the treatments within each block. A randomized block design that uses blocks with two experimental units is called a matched pairs design. Blocking helps us estimate the effects of the treatments more precisely because we can account for the variability introduced by the variables used to form the blocks.

Section 4.3: Using Studies Wisely

In this section, you learned that the types of conclusions we can draw depend on how the data are produced. When samples are selected at random, we can make inferences about the population from which the sample was drawn. However, the estimates we calculate from sample data rarely equal the true population value because of sampling variability. We can reduce sampling variability by increasing the sample size.

When treatments are applied to groups formed at random in an experiment, we can make an inference about cause and effect. Making a cause-and-effect conclusion is often difficult because it is impossible or unethical to perform certain types of experiments. Good data ethics requires that studies should be approved by an institutional review board, subjects should give informed consent, and individual data must be kept confidential.

Finally, the results of a study are statistically significant if they are too unusual to occur by chance alone.

What Did You Learn?

Learning Target	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Identify the population and sample in a statistical study.	4.1	221	R4.1
Identify voluntary response sampling and convenience sampling and explain how these sampling methods can lead to bias.	4.1	224	R4.2
Describe how to select a simple random sample with technology or a table of random digits.	4.1	228	R4.2
Describe how to select a sample using stratified random sampling and cluster sampling, distinguish stratified random sampling from cluster sampling, and give an advantage of each method.	4.1	231	R4.3
Explain how undercoverage, nonresponse, question wording, and other aspects of a sample survey can lead to bias.	4.1	234	R4.4

Learning Target	Section	Related Example on Page(s)	Relevant Chapter Review Exercise(s)
Explain the concept of confounding and how it limits the ability to make cause-and-effect conclusions.	4.2	243	R4.5
Distinguish between an observational study and an experiment, and identify the explanatory and response variables in each type of study.	4.2	244	R4.5
Identify the experimental units and treatments in an experiment.	4.2	246	R4.6
Describe the placebo effect and the purpose of blinding in an experiment.	4.2	250	R4.8
Describe how to randomly assign treatments in an experiment using slips of paper, technology, or a table of random digits.	4.2	251	R4.9
Explain the purpose of comparison, random assignment, control, and replication in an experiment.	4.2	254	R4.6, R4.8
Describe a completely randomized design for an experiment.	4.2	256	R4.6, R4.9
Describe a randomized block design and a matched pairs design for an experiment and explain the purpose of blocking in an experiment.	4.2	259, 260	R4.6, R4.9
Explain the concept of sampling variability when making an inference about a population and how sample size affects sampling variability.	4.3	271	R4.1
Explain the meaning of statistically significant in the context of an experiment and use simulation to determine if the results of an experiment are statistically significant.	4.3	274	R4.8
Identify when it is appropriate to make an inference about a population and when it is appropriate to make an inference about cause and effect.	4.3	276	R4.7
Evaluate if a statistical study has been carried out in an ethical manner.*	4.3	278	R4.10

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

Chapter 4 Review Exercises

R4.1 Nurses are the best A recent random sample of $n = 805$ adult U.S. residents found that the proportion who rated the honesty and ethical standards of nurses as high or very high is 0.85. This is 0.15 higher than the proportion recorded for doctors, the next highest-ranked profession.⁷³

- Identify the sample and the population in this setting.
- Do you think that the proportion of all U.S. residents who would rate the honesty and ethical standards of nurses as high or very high is exactly 0.85? Explain your answer.
- What is the benefit of increasing the sample size in this context?

R4.2 Parking problems The administration at a high school with 1800 students wants to gather student opinion about parking for students on campus. It isn't practical to contact all students.

- Give an example of a way to choose a voluntary response sample of students. Explain how this method could lead to bias.
- Give an example of a way to choose a convenience sample of students. Explain how this method could lead to bias.
- Describe how to select an SRS of 50 students from the school.
- Explain how the method you described in part (c) avoids the biases you described in parts (a) and (b).

R4.3 Surveying NBA fans 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 about 10% of the fans in attendance. Ticket prices for the game vary a great deal: seats near the court cost over \$200 each, while seats in the top rows

of the arena cost \$50 each. The arena is divided into 50 numbered sections, from 101 to 150. Each section has rows of seats labeled with letters from A (nearest the court) to ZZ (top row of the arena).

- Explain why it might be difficult to give the survey to an SRS of fans.
- Explain why it would be better to select a stratified random sample using the lettered rows rather than the numbered sections as strata. What is the benefit of using a stratified sample in this context?
- Explain how to select a cluster sample of fans. What is the benefit of using a cluster sample in this context?

R4.4 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, “Box office revenues are at an all-time high. How many movies have you watched in a movie theater in the past 12 months?” In all, 1131 people responded. The researchers used the responses to estimate the mean number of movies adults had watched in a movie theater over the past 12 months.

- Describe a potential source of bias related to the wording of the question. Suggest a change that would help fix this problem.
- Describe how using only residential phone numbers might lead to bias and how this will affect the estimate.
- Describe how nonresponse might lead to bias and how this will affect the estimate.

R4.5 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:⁷⁴

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 more 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.
- Identify the explanatory and response variables in this study.
- When the study looked at other variables that are related to a doctor’s choice of anesthetic, it found that Anesthetic C was not causing extra deaths. Explain the concept of confounding in this context and identify a variable that might be confounded with the doctor’s choice of anesthetic.

R4.6 Ugly fries Few people want to eat discolored french fries. To prevent spoiling and to preserve flavor, potatoes are kept refrigerated before being cut for french fries. 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, the factors, the number of levels for each factor, and the treatments.
- Describe a completely randomized design for this experiment using 300 potatoes.
- A single supplier has made 300 potatoes available to the researchers. Describe a statistical benefit and a statistical drawback of using potatoes from only one supplier.
- The researchers decided to do a follow-up experiment using potatoes from several different suppliers. Describe how they should change the design of the experiment to account for the addition of other suppliers.

R4.7 Don’t catch a cold! A recent study of 1000 students at the University of Michigan investigated how to prevent catching the common cold. The students were randomly assigned to three different cold prevention methods for 6 weeks. Some wore masks, some wore masks and used hand sanitizer, and others took no precautions. The two groups who used masks reported 10–50% fewer cold symptoms than those who did not wear a mask.⁷⁵

- Does this study allow for inference about a population? Explain your answer.
- Does this study allow for inference about cause and effect? Explain your answer.

R4.8 An herb for depression? Does the herb St. John’s wort relieve major depression? Here is an excerpt from the report of one study of this issue: “Design: Randomized, Double-Blind, Placebo-Controlled Clinical Trial.”⁷⁶ The study concluded that the difference in effectiveness of St. John’s wort and a placebo was not statistically significant.

- Describe the placebo effect in this context. How did the design of this experiment account for the placebo effect?
- Explain the purpose of the random assignment.
- Why is a double-blind design a good idea in this setting?
- Explain what “not statistically significant” means in this context.

R4.9 How long did I work? A psychologist wants to know if the difficulty of a task influences our estimate of how long we spend working at it. She designs two sets of mazes that subjects can work through on a computer. One set has easy mazes and the other has difficult mazes. Subjects work until told to stop (after 6 minutes, but subjects do not know this). They are then asked to estimate how long they worked. The psychologist has 30 students available to serve as subjects.

- (a) Describe an experiment using a completely randomized design to learn the effect of difficulty on estimated time. Make sure to carefully explain your method of assigning treatments.
- (b) Describe a matched pairs experimental design using the same 30 subjects.
- (c) Which design would be more likely to detect a difference in the effects of the treatments? Explain your answer.

R4.10* Deceiving subjects Students sign up to be subjects in a psychology experiment. When they arrive, they are told that interviews are running late and are taken to a waiting room. The experimenters then stage the theft of a valuable object left in the waiting room. Some subjects are alone with the thief, and others are present in pairs—these are the treatments being compared. Will the subject report the theft?

- (a) The students had agreed to take part in an unspecified study, and the true nature of the experiment is explained to them afterward. Does this meet the requirement of informed consent? Explain your answer.
- (b) What two other ethical principles should be followed in this study?

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

Chapter 4 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T4.1 When we take a census, we attempt to collect data from

- (a) a stratified random sample.
- (b) every individual chosen in a simple random sample.
- (c) every individual in the population.
- (d) a voluntary response sample.
- (e) a convenience sample.

T4.2 You want to take a simple random sample (SRS) of 50 of the 816 students who live in a dormitory on campus. You label the students 001 to 816 in alphabetical order. In the table of random digits, you read the entries

95592	94007	69769	33547	72450	16632	81194	14873
-------	-------	-------	-------	-------	-------	-------	-------

The first three students in your sample have labels

- (a) 955, 929, 400.
- (b) 400, 769, 769.
- (c) 559, 294, 007.
- (d) 929, 400, 769.
- (e) 400, 769, 335.

T4.3 A study of treatments for angina (pain due to low blood supply to the heart) compared bypass surgery, angioplasty, and use of drugs. The study looked at the medical records of thousands of angina patients whose doctors had chosen one of these treatments. It found that the average survival time of patients given drugs was the highest. What do you conclude?

- (a) This study proves that drugs prolong life and should be the treatment of choice.
- (b) We can conclude that drugs prolong life because the study was a comparative experiment.
- (c) We can't conclude that drugs prolong life because the patients were volunteers.
- (d) We can't conclude that drugs prolong life because the groups might differ in ways besides the treatment.
- (e) We can't conclude that drugs prolong life because no placebo was used.

T4.4 Tonya wanted to estimate the average amount of time that students at her school spend on Facebook each day. She gets an alphabetical roster of students in the school from the registrar's office and numbers the students from 1 to 1137. Then Tonya uses a random number generator to pick 30 distinct labels from 1 to 1137. She surveys those 30 students about their Facebook use. Tonya's sample is a simple random sample because

- (a) it was selected using a chance process.
- (b) it gave every individual the same chance to be selected.
- (c) it gave every possible sample of size 30 an equal chance to be selected.
- (d) it doesn't involve strata or clusters.
- (e) it is guaranteed to be representative of the population.

T4.5 Consider an experiment to investigate the effectiveness of different insecticides in controlling pests and their impact on the productivity of tomato plants. What is the best reason for randomly assigning treatment levels (spraying or not spraying) to the experimental units (farms)?

- (a) Random assignment eliminates the effects of other variables, like soil fertility.
- (b) Random assignment eliminates chance variation in the responses.
- (c) Random assignment allows researchers to generalize conclusions about the effectiveness of the insecticides to all farms.
- (d) Random assignment will tend to average out all other uncontrolled factors such as soil fertility so that they are not confounded with the treatment effects.
- (e) Random assignment helps avoid bias due to the placebo effect.

T4.6 Researchers randomly selected 1700 people from Canada who had never suffered a heart attack and rated the happiness of each person. Ten years later, the researchers followed up with each person and found that people who were initially rated as happy were less likely to have a heart problem.⁷⁷ Which of the following is the most appropriate conclusion based on this study?

- (a) Happiness causes better heart health for all people.
- (b) Happiness causes better heart health for Canadians.
- (c) Happiness causes better heart health for the 1700 people in the study.
- (d) Happier people in Canada are less likely to have heart problems.
- (e) Happier people in the study were less likely to have heart problems.

T4.7 A TV station wishes to obtain information on the TV viewing habits in its market area. The market area contains one city of population 170,000, another city of 70,000, and four towns of about 5000 residents each. The station suspects that the viewing habits may be different in larger and smaller cities and in the rural areas. Which of the following sampling designs would yield the type of information the station requires?

- (a) A stratified sample from the cities and towns in the market area
- (b) A cluster sample using the cities and towns as clusters
- (c) A convenience sample from the market area
- (d) A simple random sample from the market area
- (e) An online poll that invites all people from the cities and towns in the market area to participate

T4.8 Bias in a sampling method is

- (a) any difference between the sample result and the truth about the population.
- (b) the difference between the sample result and the truth about the population due to using chance to select a sample.
- (c) any difference between the sample result and the truth about the population due to practical difficulties such as contacting the subjects selected.
- (d) any difference between the sample result and the truth about the population that tends to occur in the same direction whenever you use this sampling method.
- (e) racism or sexism on the part of those who take the sample.

T4.9 You wonder if TV ads are more effective when they are longer or repeated more often or both. So you design an experiment. You prepare 30-second and 60-second ads for a camera. Your subjects all watch the same TV program, but you assign them at random to four groups. One group sees the 30-second ad once during the program; another sees it three times; the third group sees the 60-second ad once; and the last group sees the 60-second ad three times. You ask all subjects how likely they are to buy the camera. Which of the following best describes the design of this experiment?

- (a) This is a randomized block design, but not a matched pairs design.
- (b) This is a matched pairs design.
- (c) This is a completely randomized design with one explanatory variable (factor).
- (d) This is a completely randomized design with two explanatory variables (factors).
- (e) This is a completely randomized design with four explanatory variables (factors).

T4.10 Can texting make you healthier? Researchers randomly assigned 700 Australian adults to either receive usual health care or usual health care plus automated text messages with positive messages, such as "Walking is cheap. It can be done almost anywhere. All you need is comfortable shoes and clothing." The group that received the text messages showed a statistically significant increase in physical activity.⁷⁸ What is the meaning of "statistically significant" in this context?

- (a) The results of this study are very important.
- (b) The results of this study should be generalized to all people.
- (c) The difference in physical activity for the two groups is greater than 0.

- (d) The difference in physical activity for the two groups is very large.
- (e) The difference in physical activity for the two groups is larger than the difference that could be expected to happen by chance alone.

T4.11 You want to know the opinions of American high school teachers on the issue of establishing a national proficiency test as a prerequisite for graduation from high school. You obtain a list of all high school teachers belonging to the National Education Association (the country's largest teachers' union) and mail a survey to a random sample of 2500 teachers. In all, 1347 of the teachers return the survey. Of those who responded, 32% say that they favor some kind of national proficiency test. Which of the following statements about this situation is true?

- (a) Because random sampling was used, we can feel confident that the percent of all American high school teachers who would say they favor a national proficiency test is close to 32%.
- (b) We cannot trust these results, because the survey was mailed. Only survey results from face-to-face interviews are considered valid.
- (c) Because over half of those who were mailed the survey actually responded, we can feel fairly confident that the actual percent of all American high school teachers who would say they favor a national proficiency test is close to 32%.
- (d) The results of this survey may be affected by undercoverage and nonresponse.
- (e) The results of this survey cannot be trusted due to voluntary response bias.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T4.12 Elephants sometimes damage trees in Africa. It turns out that elephants dislike bees. They recognize beehives in areas where they are common and avoid them. Can this information be used to keep elephants away from trees? Researchers want to design an experiment to answer these questions using 72 acacia trees and three treatments: active hives, empty hives, and no hives.⁷⁹

- (a) Identify the experimental units in this experiment.
- (b) Explain why it is beneficial to include some trees that have no hives.
- (c) Describe how the researchers could carry out a completely randomized design for this experiment. Include a description of how the treatments should be assigned.

T4.13 Google and Gallup teamed up to survey a random sample of 1673 U.S. students in grades 7–12. One of the questions was “How confident are you that you could learn computer science if you wanted to?” Overall, 54% of students said they were very confident.⁸⁰

- (a) Identify the population and the sample.
- (b) Explain why it was better to randomly select the students rather than putting the survey question on a website and inviting students to answer the question.
- (c) Do you expect that the percent of all U.S. students in grades 7–12 who would say “very confident” is exactly 54%? Explain your answer.
- (d) The report also broke the results down by gender. For this question, 62% of males and 48% of females said they were very confident. Which of the three estimates (54%, 62%, 48%) do you expect is closest to the value it is trying to estimate? Explain your answer.

T4.14 Many people start their day with a jolt of caffeine from coffee or a soft drink. Most experts agree that people who consume large amounts of caffeine each day may suffer from physical withdrawal symptoms if they stop ingesting their usual amounts of caffeine. Researchers recruited 11 volunteers who were caffeine dependent and who were willing to take part in a caffeine withdrawal experiment. The experiment was conducted on two 2-day periods that occurred one week apart. During one of the 2-day periods, each subject was given a capsule containing the amount of caffeine normally ingested by that subject in one day. During the other study period, the subjects were given placebos. The order in which each subject received the two types of capsules was randomized. The subjects' diets were restricted during each of the study periods. At the end of each 2-day study period, subjects were evaluated using a tapping task in which they were instructed to press a button 200 times as fast as they could.⁸¹

- (a) Identify the explanatory and response variables in this experiment.
- (b) How was blocking used in the design of this experiment? What is the benefit of blocking in this context?
- (c) Researchers randomized the order of the treatments to avoid confounding. Explain how confounding might occur if the researchers gave all subjects the placebo first and the caffeine second. In this context, what problem does confounding cause?
- (d) Could this experiment have been carried out in a double-blind manner? Explain your answer.

Chapter 4 Project Response Bias

In this project, your team will design and conduct an experiment to investigate the effects of response bias in surveys.⁸² 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.

- Does the wording of a question affect the response?
- Do the characteristics of the interviewer affect the response?
- Does anonymity change the responses to sensitive questions?
- Does manipulating the answer choices affect the response?
- Can revealing other peoples' answers to a question change the response?

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

- (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) An explanation of the treatments in your experiment and how you will determine which subjects get which treatment.

- (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 experiments.

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

- i. "Do you find females who wear makeup attractive?" (Questioner wearing makeup: 75% answered "Yes.")
- ii. "Do you find females who wear makeup attractive?" (Questioner not wearing makeup: 30% answered "Yes.")

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

- i. "Do you watch cartoons?" (90% answered "Yes.")
 - ii. "Do you *still* watch cartoons?" (60% answered "Yes.")
2. Once your teacher has approved your design, carry out the experiment. Record your data in a table.
 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.

Cumulative AP[®] Practice Test 1

Section I: Multiple Choice Choose the best answer for Questions AP1.1–AP1.14.

AP1.1 You look at real estate ads for houses in Sarasota, Florida. Many houses have prices from \$200,000 to \$400,000. The few houses on the water, however, have prices up to \$15 million. Which of the following statements best describes the distribution of home prices in Sarasota?

- (a) The distribution is most likely skewed to the left, and the mean is greater than the median.
- (b) The distribution is most likely skewed to the left, and the mean is less than the median.
- (c) The distribution is roughly symmetric with a few high outliers, and the mean is approximately equal to the median.
- (d) The distribution is most likely skewed to the right, and the mean is greater than the median.
- (e) The distribution is most likely skewed to the right, and the mean is less than the median.

AP1.2 A child is 40 inches tall, which places her at the 90th percentile of all children of similar age. The heights for children of this age form an approximately Normal distribution with a mean of 38 inches. Based on this information, what is the standard deviation of the heights of all children of this age?

- (a) 0.20 inch
- (b) 0.31 inch
- (c) 0.65 inch
- (d) 1.21 inches
- (e) 1.56 inches

AP1.3 A large set of test scores has mean 60 and standard deviation 18. If each score is doubled, and then 5 is subtracted from the result, the mean and standard deviation of the new scores are

- (a) mean 115 and standard deviation 31.
- (b) mean 115 and standard deviation 36.
- (c) mean 120 and standard deviation 6.
- (d) mean 120 and standard deviation 31.
- (e) mean 120 and standard deviation 36.

AP1.4 For a certain experiment, the available experimental units are eight rats, of which four are female (F1, F2, F3, F4) and four are male (M1, M2, M3, M4). There are to be four treatment groups, A, B, C, and D. If a randomized block design is used, with the experimental units blocked by gender, which of the following assignments of treatments is impossible?

- (a) $A \rightarrow (F1, M1), B \rightarrow (F2, M2), C \rightarrow (F3, M3), D \rightarrow (F4, M4)$
- (b) $A \rightarrow (F1, M2), B \rightarrow (F2, M3), C \rightarrow (F3, M4), D \rightarrow (F4, M1)$
- (c) $A \rightarrow (F1, M2), B \rightarrow (F3, F2), C \rightarrow (F4, M1), D \rightarrow (M3, M4)$
- (d) $A \rightarrow (F4, M1), B \rightarrow (F2, M3), C \rightarrow (F3, M2), D \rightarrow (F1, M4)$
- (e) $A \rightarrow (F4, M1), B \rightarrow (F1, M4), C \rightarrow (F3, M2), D \rightarrow (F2, M3)$

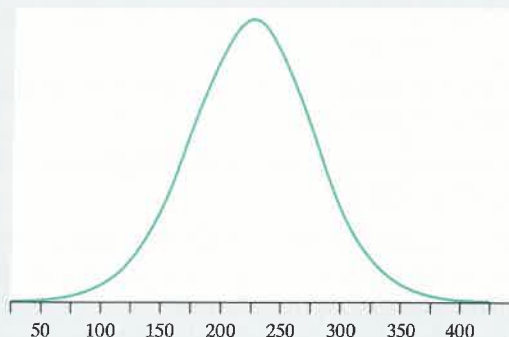
AP1.5 For a biology project, you measure the weight in grams (g) and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is

$$\text{predicted tail length} = 20 + 3 \times \text{weight}$$

Which of the following is *not* correct?

- (a) The slope is 3, which indicates that a mouse's predicted tail length should increase by about 3 mm for each additional gram of weight.
- (b) The predicted tail length of a mouse that weighs 38 grams is 134 millimeters.
- (c) By looking at the equation of the least-squares line, you can see that the correlation between weight and tail length is positive.
- (d) If you had measured the tail length in centimeters instead of millimeters, the slope of the regression line would have been $3/10 = 0.3$.
- (e) Mice that have a weight of 0 grams will have a tail of length 20 mm.

AP1.6 The figure shows a Normal density curve. Which of the following gives the best estimates for the mean and standard deviation of this Normal distribution?



- (a) $\mu = 200, \sigma = 50$
- (b) $\mu = 200, \sigma = 25$
- (c) $\mu = 225, \sigma = 50$
- (d) $\mu = 225, \sigma = 25$
- (e) $\mu = 225, \sigma = 275$

AP1.7 The owner of a chain of supermarkets notices that there is a positive correlation between the sales of beer and the sales of ice cream over the course of the previous year. During seasons when sales of beer were above average, sales of ice cream also tended to be above average. Likewise, during seasons when sales of beer were below average, sales of ice cream also tended to be below average. Which of the following would be a valid conclusion from these facts?

- (a) Sales records must be in error. There should be no association between beer and ice cream sales.
- (b) Evidently, for a significant proportion of customers of these supermarkets, drinking beer causes a desire for ice cream or eating ice cream causes a thirst for beer.
- (c) A scatterplot of monthly ice cream sales versus monthly beer sales would show that a straight line describes the pattern in the plot, but it would have to be a horizontal line.
- (d) There is a clear negative association between beer sales and ice cream sales.
- (e) The positive correlation is most likely a result of the variable temperature; that is, as temperatures increase, so do both beer sales and ice cream sales.

AP1.8 Here are the IQ scores of 10 randomly chosen fifth-grade students:

145	139	126	122	125	130	96	110	118	118
-----	-----	-----	-----	-----	-----	----	-----	-----	-----

Which of the following statements about this data set is *not* true?

- (a) The student with an IQ of 96 is considered an outlier by the $1.5 \times \text{IQR}$ rule.
- (b) The five-number summary of the 10 IQ scores is 96, 118, 123.5, 130, 145.
- (c) If the value 96 were removed from the data set, the mean of the remaining 9 IQ scores would be greater than the mean of all 10 IQ scores.
- (d) If the value 96 were removed from the data set, the standard deviation of the remaining 9 IQ scores would be less than the standard deviation of all 10 IQ scores.
- (e) If the value 96 were removed from the data set, the IQR of the remaining 9 IQ scores would be less than the IQR of all 10 IQ scores.

AP1.9 Before he goes to bed each night, Mr. Kleen pours dishwasher powder into his dishwasher and turns it on. Each morning, Mrs. Kleen weighs the box of dishwasher powder. From an examination of the data, she concludes that Mr. Kleen dispenses a rather consistent amount of powder each night. Which of the following statements is true?

- I. There is a high positive correlation between the number of days that have passed since the box of dishwasher powder was opened and the amount of powder left in the box.
 - II. A scatterplot with days since purchase as the explanatory variable and total amount of dishwasher powder used as the response variable would display a strong positive association.
 - III. The correlation between the amount of powder left in the box and the amount of powder used should be -1 .
- (a) I only
 - (b) II only
 - (c) III only
 - (d) II and III only
 - (e) I, II, and III

AP1.10 The General Social Survey (GSS), conducted by the National Opinion Research Center at the University of Chicago, is a major source of data on social attitudes in the United States. Once each year, 1500 adults are interviewed in their homes all across the country. The subjects are asked their opinions about sex and marriage; attitudes toward

women, welfare, foreign policy; and many other issues. The GSS begins by selecting a sample of counties from the 3000 counties in the country. The counties are divided into urban, rural, and suburban; a separate sample of counties is chosen at random from each group. This is a

- (a) simple random sample.
- (b) systematic random sample.
- (c) cluster sample.
- (d) stratified random sample.
- (e) voluntary response sample.

AP1.11 You are planning an experiment to determine the effect of the brand of gasoline and the weight of a car on gas mileage measured in miles per gallon. You will use a single test car, adding weights so that its total weight is 3000, 3500, or 4000 pounds. The car will drive on a test track at each weight using each of Amoco, Marathon, and Speedway gasoline. Which is the best way to organize the study?

- (a) Start with 3000 pounds and Amoco and run the car on the test track. Then do 3500 and 4000 pounds. Change to Marathon and go through the three weights in order. Then change to Speedway and do the three weights in order once more.
- (b) Start with 3000 pounds and Amoco and run the car on the test track. Then change to Marathon and then to Speedway without changing the weight. Then add weights to get 3500 pounds and go through the three gasolines in the same order. Then change to 4000 pounds and do the three gasolines in order again.
- (c) Choose a gasoline at random, and run the car with this gasoline at 3000, 3500, and 4000 pounds in order. Choose one of the two remaining gasolines at random and again run the car at 3000, then 3500, then 4000 pounds. Do the same with the last gasoline.
- (d) There are nine combinations of weight and gasoline. Run the car several times using each of these combinations. Make all these runs in random order.
- (e) Randomly select an amount of weight and a brand of gasoline, and run the car on the test track. Repeat this process a total of 30 times.

AP1.12 A linear regression was performed using the following five data points: A(2, 22), B(10, 4), C(6, 14), D(14, 2), E(18, -4). The residual for which of the five points has the largest absolute value?

- (a) A
- (b) B
- (c) C
- (d) D
- (e) E

AP1.13 The frequency table summarizes the distribution of time that 140 patients at the emergency room of a small-city hospital waited to receive medical attention during the last month.

Waiting time	Frequency
Less than 10 minutes	5
At least 10 but less than 20 minutes	24
At least 20 but less than 30 minutes	45
At least 30 but less than 40 minutes	38
At least 40 but less than 50 minutes	19
At least 50 but less than 60 minutes	7
At least 60 but less than 70 minutes	2

Which of the following represents possible values for the median and *IQR* of waiting times for the emergency room last month?

- (a) median = 27 minutes and *IQR* = 15 minutes
- (b) median = 28 minutes and *IQR* = 25 minutes
- (c) median = 31 minutes and *IQR* = 35 minutes
- (d) median = 35 minutes and *IQR* = 45 minutes
- (e) median = 45 minutes and *IQR* = 55 minutes

Section II: Free Response Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

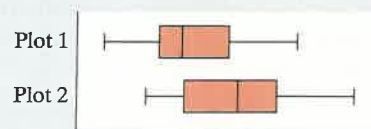
AP1.15 The manufacturer of exercise machines for fitness centers has designed two new elliptical machines that are meant to increase cardiovascular fitness. The two machines are being tested on 30 volunteers at a fitness center near the company's headquarters. The volunteers are randomly assigned to one of the machines and use it daily for two months. A measure of cardiovascular fitness is administered at the start of the experiment and again at the end. The following stemplot contains the differences (After – Before) in the two scores for the two machines. Note that higher scores indicate larger gains in fitness.

Machine A		Machine B
	0	2
54	1	0
876320	2	159
97411	3	2489
61	4	257
	5	359

Key: 2 | 1 represents a difference (After – Before) of 21 in fitness scores.

- (a) Write a few sentences comparing the distributions of cardiovascular fitness gains from the two elliptical machines.
- (b) Which machine should be chosen if the company wants to advertise it as achieving the highest overall gain in cardiovascular fitness? Explain your reasoning.

AP1.14 Boxplots of two data sets are shown.



Based on the boxplots, which of the following is true?

- (a) The range of both plots is about the same.
- (b) The means of both plots are approximately equal.
- (c) Plot 2 contains more data points than Plot 1.
- (d) The medians are approximately equal.
- (e) Plot 1 is more symmetric than Plot 2.

- (c) Which machine should be chosen if the company wants to advertise it as achieving the most consistent gain in cardiovascular fitness? Explain your reasoning.
- (d) Give one reason why the advertising claims of the company (the scope of inference) for this experiment would be limited. Explain how the company could broaden that scope of inference.

AP1.16 Those who advocate for monetary incentives in a work environment claim that this type of incentive has the greatest appeal because it allows the winners to do what they want with their winnings. Those in favor of tangible incentives argue that money lacks the emotional appeal of, say, a weekend for two at a romantic country inn or elegant hotel, or a weeklong trip to Europe.

A few years ago a national tire company, in an effort to improve sales of a new line of tires, decided to test which method—offering cash incentives or offering non-cash prizes such as vacations—was more successful in increasing sales. The company had 60 retail sales districts of various sizes across the country and data on the previous sales volume for each district.

- (a) Describe a completely randomized design using the 60 retail sales districts that would help answer this question.

- (b) Explain how you would use the following excerpt from the table of random digits to do the random assignment that your design requires. Then use your method to make the first three assignments.

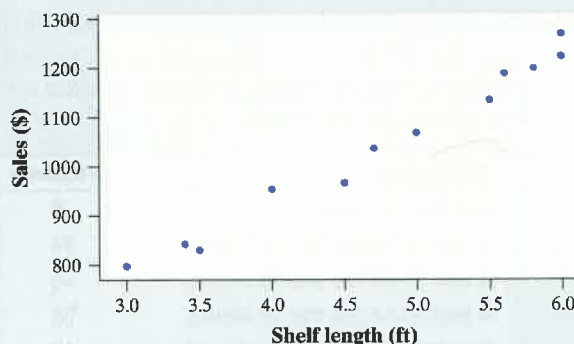
07511	88915	41267	16853	84569	79367	32337	03316
81486	69487	60513	09297	00412	71238	27649	39950

- (c) One of the company's officers suggested that it would be better to use a matched pairs design instead of a completely randomized design. Explain how you would change your design to accomplish this.

AP1.17 In retail stores, there is a lot of competition for shelf space. There are national brands for most products, and many stores carry their own line of in-house brands, too. Because shelf space is not infinite, the question is how many linear feet to allocate to each product and which shelf (top, bottom, or somewhere in the middle) to put it on. The middle shelf is the most popular and lucrative, because many shoppers, if undecided, will simply pick the product that is at eye level.

A local store that sells many upscale goods is trying to determine how much shelf space to allocate to its own brand of men's personal-grooming products. The middle shelf space is randomly varied between 3 and 6 linear feet over the next 12 weeks, and weekly sales revenue (in dollars) from the store's brand of personal-grooming products for men is recorded. Here is some computer output from the study, along with a scatterplot:

Predictor	Coef	SE Coef	T	P
Constant	317.940	31.32	10.15	0.000
Shelf length	152.680	6.445	23.69	0.000
S = 22.9212 R-Sq = 98.2% R-Sq(adj) = 98.1%				



- Describe the relationship between shelf length and sales.
- Write the equation of the least-squares regression line. Be sure to define any variables you use.
- If the store manager were to decide to allocate 5 linear feet of shelf space to the store's brand of men's grooming products, what is the best estimate of the weekly sales revenue?
- Interpret the value of s .
- Identify and interpret the coefficient of determination.

AP1.18 The manager of the store in the preceding exercise calculated the residual for each point in the scatterplot and made a dotplot of the residuals. The distribution of residuals is roughly Normal with a mean of \$0 and standard deviation of \$22.92.

- What percent of the actual sales amounts do you expect to be within \$5 of their expected sales amount?
- The middle 95% of residuals should be between which two values? Use this information to give an interval of plausible values for the weekly sales revenue if 5 linear feet are allocated to the store's brand of men's grooming products.