

SECTION 2.

Variation Out of Control

How can you do your best at your next midterm? Would taking a nap (before, not during, the exam) help? How could you find out? You could just try it, of course, but then if you did really well, you wouldn't know whether it was because of the nap or because it was just an easy exam for you. You could try it a few times, sometimes napping and sometimes not, but there's a lot of variability involved here. Sometimes your naps will be longer or your sleep deeper. Sometimes the exams will be harder or easier, and your level of preparation for the material will vary. And even if napping works for you, how could you know whether it would work for someone else? This natural variability, both between people and between contexts, will make it difficult to determine just what effect, if any, napping has on test performance.

For another example, consider the need to test the effectiveness of vaccines. Beginning in mid-2009, many people began to worry about the flu, as news of the H1N1 virus (also known as the "swine flu" virus) swept the world. Eventually, a vaccine was made available, but this raised more questions. Was the vaccine really effective? Were two doses required, or was one sufficient? Could the vaccine be inhaled, or was it better to have it injected? Certainly you could get the vaccine yourself and then see whether you got the flu, but if you remained healthy, it might be just because you weren't exposed to the virus. Or if you got sick, it might be just because, even with the vaccine, your body was susceptible to this particular virus. Again, natural variability makes it difficult to determine how effective a vaccine, or a particular vaccine regimen, is at preventing the flu.

Statisticians and researchers have developed several methods for controlling variability in order to find answers to important questions like these. In Chapter 1 we explained why these questions can be answered only with controlled experiments. In this chapter, we show you how it's done.

Review of Experimental Basics

Before getting to the heart of the chapter, let's review a few of the important concepts covered in Chapter 1.

The phrase *cause and effect* has a common English meaning, but nonetheless we need to be clear about its meaning in this context. For our purposes, we say that two variables X and Y have a cause-and-effect relationship if, whenever we change X for a person (or object), then Y typically changes in a predictable way. For example, if napping causes test scores to improve, then if you changed your test preparation from "no-nap" to "nap," you would see an increase in your test scores. If we gave you a flu vaccine, we would change your status from "not vaccinated" to "vaccinated," and we would hope to see your risk of getting the flu go down as a result.

We call the variable that we change—the "cause"—the **treatment variable**. The variable that we are interested in seeing changed—the "effect"—is called the **response variable**. The treatment variable records, for example, whether a subject received the vaccine or did not get the vaccine, and the response variable records whether or not that person later got the flu.

The trick is to determine whether changes that we see in the response variable are due to changes in the treatment variable; are due to changes in some other, unseen variables; or are due to chance. Unseen variables that might affect the response vari-

For example, in an evaluation of an SAT preparation company, researchers found that students who completed the SAT preparation course did better on the SAT than students who did not. The treatment variable here records whether or not a student took the preparation class, and the response variable records the student's SAT scores. One problem with the study was that the preparation course was expensive, and a student's decision to take the course depended somewhat on the student's ability to pay (or the family's ability to pay)—a possible confounding variable.

To be a confounding variable, a variable must affect *both* the treatment *and* the response. The student's ability to pay meets this criterion, because it affects both the treatment (attend or not-attend the SAT preparation course) and the response (score on SAT). A student's ability to pay for SAT preparation affects the treatment variable because, on average, students who take the SAT preparation class are better off financially than students in the group who do not take the class. And this ability to pay means that the students might, typically, have increased access to other resources that help them do better on the SAT (for example, they might not need jobs and so would have more time to prepare); for this reason, the response variable is also affected.

Another possible confounding variable is motivation. Students who are motivated enough to spend money on a prep course, to go through the trouble of finding a prep course, and to put in the extra time attending the course might also be motivated enough to study in other ways. So even if the prep course is not effective, the students' extra motivation might be the reason for their higher SAT scores (Briggs 2002).

Maybe you can think of other confounding variables as well. The presence of confounding variables means we cannot conclude that the prep course caused the higher SAT scores. To be able to conclude that the prep course caused the change, we would have to eliminate all other explanations.

To eliminate other explanations, the experimenter must make sure that the subjects in the experiment are alike in every way except that they have different values for the treatment variable. To achieve this, experimenters can take control and assign subjects to a treatment group. When this is done, we call the experiment a **controlled experiment**.

Still, making sure that the subjects in the different treatment groups are all alike is difficult, because people (and animals and even some inanimate objects) are not all alike. This variability means that if we're not careful, some differences between the groups will arise that might be confounding variables and prevent us from concluding that the treatment caused any observed changes in the response variable.

Researchers use **random assignment** to ensure that different treatment groups are as similar as possible. To take a simple example, if we have two treatment groups, one getting a vaccine and the other a placebo, then by randomly assigning subjects to the two groups, we will have roughly equal proportions of men and women in both groups. In that way, gender cannot be a confounding factor, because both groups will have about the same numbers of each gender. Random assignment ensures that other variables, even some we don't know about, will be distributed in roughly equal proportions among the groups.

Observational studies are quite common, in part because it is not always possible to do a controlled experiment. How could researchers decide whether using a cell phone over many years causes brain cancer in a randomized study? First, such a study would be unethical. Second, even if they ignored ethical considerations, how could researchers make sure that subjects who were assigned to the treatment group used their phones often enough over many years? And how could they be sure that subjects in the control group did *not* use cell phones for many years?

A major drawback with observational studies is that they can only establish an association between the treatment and response variables; they cannot establish a cause-and-effect relationship. The news media often ignore this fact and report the



Looking Back

Observational Studies and Confounding

Recall that in an observational study, we can never eliminate the possibility that a confounding variable exists. This means we cannot reach cause-and-effect conclusions on the basis of a single observational study.

results of observational studies as though they were controlled experiments. For example, the BBC News published this headline: “Happiness Wards Off Heart Disease, Study Suggests” (<http://news.bbc.co.uk>, viewed January 25, 2011, via <http://jonathan.mueller.faculty.noctrl.edu>). The reported study must be an observational study, because how could researchers assign people to be happy for a long period of time? Therefore, we can’t conclude that happiness causes a lower risk of heart disease, despite the claims of the headline.



KEY POINT

Random assignment to treatment groups ensures that the subjects in all groups are as similar as possible, so that the only real difference between groups is the treatment they receive.

EXAMPLE 1 Improving Memory

Which is better for improving memory: a short nap, caffeine, or nothing? Researchers at the Veterans Administration in San Diego randomly assigned 61 subjects to one of three groups. All subjects took a basic memory test. Depending on which group they were assigned to, subjects then took a 30-minute nap, or had a moderate dose of caffeine, or did nothing. Afterwards, subjects were again tested on their memory. Researchers found that the nappers improved more on the memory tests than either the caffeine takers or the placebo group (Mednick et al. 2008).

QUESTION For this controlled experiment, identify the treatment and response variables. Restate the conclusion of the study in terms of a cause-and-effect conclusion.

SOLUTION The treatment variable here records the treatment group that each subject was assigned to: subjects napped, or had caffeine, or did neither. The response variable is the change in the subjects’ memory test score. The conclusion could be stated as “Napping improves memory.”

TRY THIS! Exercise 12.9



Statistical Power

Random assignment helps researchers eliminate (or at least minimize) the effects of possible confounding variables. But there is always the possibility that observed changes in the response variable are simply due to chance, rather than due to the treatment the subjects received.

Because statistical tests involve chance, even when the two groups we’re comparing are truly different, there is a probability that we won’t see this difference. On the other hand, there is a probability that we will correctly find that the groups are truly different. This probability, called the **power**, was introduced in Chapter 8. The power is the probability that, for a given value of the parameter, we will reject the null hypothesis. In other words, power measures the probability of detecting differences that really exist between groups.

Obviously, we want this probability to be big. Many researchers like to have a power of at least 80%. Higher is even better.

The power depends primarily on three things: the sample size, the size of the true difference between the groups, and the natural variability within the population. Of these three factors, researchers and statisticians have direct control only over sample size. Still, by identifying sources of natural variability in the population, researchers can design experiments that can improve the power by controlling for the

Variability that exists within a population can arise from several different sources. One source is natural variability: the subjects (or objects) being studied naturally have variability. If we are comparing the weights of different groups of people, we should expect natural variability within these groups, because everyone's weight is slightly different. Another important source of variability is measurement error. This is the variability caused by the devices or methods used to measure something. For example, if you weigh yourself twice in a row, you don't always get exactly the same weight the second time. Figure 12.1 shows the distribution of weights of two thousand 1-euro coins. You might expect coins would weigh exactly the same, because heavier coins might be perceived as more valuable. Still, whether as a consequence of measurement error or of variations in the coin-minting process, we still see variability.

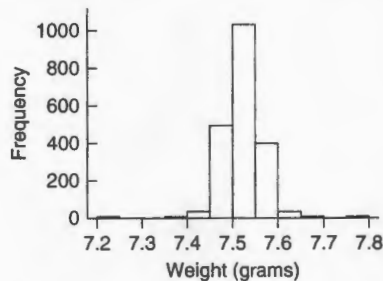


FIGURE 12.1 Even though the minting of coins is very carefully controlled, there is still variability, as this distribution of the weights of two thousand 1-euro coins shows.

The best way to control for this variation when conducting a research study is to hold a variable fixed at a particular value. This is called **controlling for a variable**. If we are performing a study to see how naps affect memory, and we believe that caffeine intake affects naps, then we might control for caffeine amounts by requiring everyone to be caffeine-free. Or we might even give all of the groups the same amount of caffeine, if we were interested in seeing how our conclusions were affected by caffeine.

Blocking

For factors that the researchers cannot control, they can apply blocking. **Blocking** is a technique in which researchers group similar objects into "blocks" and then assign treatments randomly within each block. Blocking helps achieve two purposes. It reduces bias (particularly useful when the sample sizes are small), and it increases statistical power.

In many medical studies, for example, different age groups respond differently to treatments. Researchers cannot control a person's age, but they can create blocks of subjects who are the same age (or nearly the same age) and in that way increase the statistical power of their study.

The basic idea is that once we've grouped similar subjects or objects together, we hope that subjects within a block will respond similarly to the treatment. If so, then we don't need to measure quite so many people, because there is less variability than we originally had.

KEY POINT

A block consists of subjects or objects that are similar on one or more variables. Objects *within* a block are then randomly assigned to treatment groups.

Creating Blocks

To create blocks, ask whether you know of any variables that might affect the outcome and cannot be controlled. If so, create the blocks by grouping together objects that are similar to one another on those variables.

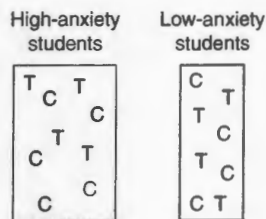


FIGURE 12.2 Each letter represents a student in the study. Students within each block are randomly assigned either to the Expressive Writing treatment (T) or to the control group (C).

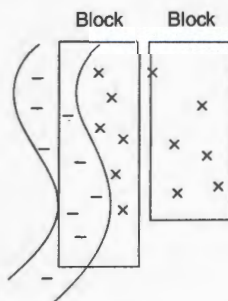


FIGURE 12.3 The X's represent the locations of yucca plants. The plants near the (usually dry) riverbed have different soil types and moisture than the other plants, so it makes sense to create blocks grouping together the plants that are in a similar environment.

Once you have created the blocks, you then randomly assign objects within each block either to the control group or to the treatment group (or to the many different treatment groups, if the study is considering several different treatments). Each block should have objects in all treatment and control groups.

For example, psychologists at the University of Chicago had evidence that “expressive writing”—writing about emotions—might help prevent students from “choking” on exams (Ramirez and Beilock 2011). To test this, they recruited ninth grade biology students at a large high school. Students were evaluated about their upcoming final exam. Some students were very anxious, but others were much less anxious. Because researchers suspected that one source of variation in the effects of the treatment might be the student’s anxiety level, they created two blocks. One block consisted of students who were the most anxious about the final exam. The other block consisted of the least anxious students. Ten minutes before the exam, students in each block were randomly assigned to do one of two activities. Half of the students were told to write about their anxieties concerning the test; this expressive writing constituted the treatment. The other half of the students were told to write about any topic that was not covered by the exam; these students formed the control group. Figure 12.2 illustrates this blocking design.

At the end of the study, researchers found that the high-anxiety students who performed expressive writing outperformed the high-anxiety students in the control group. On the other hand, low-anxiety students who did expressive writing did not perform any differently than their control-group counterparts. The blocking structure of the experiment helped the researcher tease out some complex relations among anxiety, the treatment, and exam performance.

Blocking is often used in agricultural and biological studies done “in the field.” The reason is that when you are studying plants in the great outdoors, the amount of sunlight, water, or soil nutrients that a plant might receive can vary widely just a few feet in any direction. Figure 12.3 shows an illustration of two blocks used to study the growth of Mojave Yucca. Researchers, concerned about campers and outdoor recreationists damaging the plants, are wondering whether the plants can be made healthier by protecting some with small fences.

The illustration shows that some of the plants are near a riverbed, which fills with water after a rainstorm. Plants near the riverbed receive more moisture, so if a study is being done on these plants, then the plants near the river should be put in one block, and the other plants in another. Then, within each block, some plants are randomly chosen to be surrounded by fences, and others are randomly chosen to have no protection.

EXAMPLE 2 Wash That Statistician Right Out of Your Hair

A cosmetics company has created a new dandruff shampoo, which it believes will work better than its older shampoo formula. Company investigators believe that the effects of the shampoo depend on hair color. For this reason, they plan to block on hair color. To do this, they create blocks that have people with blond hair, brown hair, black hair, red hair, and grey hair, respectively. They then randomly select three of the blocks to receive the new shampoo and the other two blocks to receive the old shampoo.

QUESTION Is this an effective design for the study? If not, describe an improvement.

SOLUTION No. The researchers randomly assigned entire blocks to treatment groups. Instead, they should randomize *within* blocks. If an entire block is assigned to the same treatment—for example, if all of the black-hair subjects are assigned to the old shampoo—then researchers won’t be able to determine whether the new shampoo works on black hair.

To properly conduct this study, they should randomly assign half of the blond-hair group to receive the new formula and half to receive the old. Then they should randomly assign half of the brown-hair subjects to receive the new formula and half to receive the old, and so on.

TRY THIS! Exercise 12.17



EXAMPLE 3 Baking Cookies

An amateur baker wants to experiment with two different cookie recipes. One recipe calls for margarine, one for butter. The problem is that the oven the baker is using is a bit old and doesn't heat up consistently. If the oven is set to 350 degrees Fahrenheit, not every region in the oven will be exactly the same temperature. Through experience, the baker has some ideas about which areas are hotter and which are cooler, and these are shown in Figure 12.4. This figure also shows where each cookie will be placed.

QUESTION: Explain how to create blocks to test which cookie recipe tastes best.

SOLUTION: The baker should create three blocks. The cookies within the 370°F zone (cookies 3, 4, and 8) should be in one block. The cookies within the 360°F zone (cookies 1, 2, 5, 6, 7, 9, 10, 11, 12, 15, and 16) are in another block, and the cookies in the 350°F zone (cookies 13, 14, 17, 18, 19, and 20) go in another block. Within each block, half of the spaces are randomly chosen to receive a margarine cookie, and the other half receive butter cookies. This can be done by putting the cookie numbers into three hats. One hat has the numbers for the 370°F block (numbers 3, 4, and 8), another hat has the numbers for the 360°F block, and the third hat has the numbers for the 350°F block. From each hat, half the numbers are drawn, and the selected numbers will be the margarine cookies. For blocks with an odd number of spaces, a coin-flip can be used to determine whether the last space gets margarine (heads) or butter (tails).



TRY THIS! Exercise 12.19

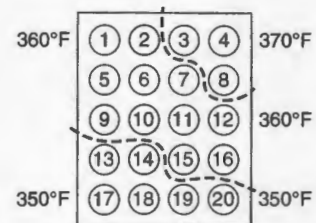


FIGURE 12.4 The oven has three temperature regions, and baking conditions are similar within each region.

Blocking is useful because if you have correctly identified a variable that leads to different outcomes in your study, then by grouping subjects together on the basis of that variable, you have decreased the variability in the study—and decreasing the variability leads to greater statistical power.

Studies with small sample sizes can be greatly improved by blocking. Suppose that you wish to replicate the study about the effects of expressive writing on students in your own class. Imagine that there are 20 students, and most—say, 15 students—are not anxious. If you do not create blocks before you randomly assign 10 students to the treatment group and the other 10 to control, then it is possible that just by chance, all 5 of the anxious students will be in the control group. If this happens, then you will not be able to see the effect of the treatment, because the people who would be best helped by the treatment will not receive it.

Instead, if you created blocks, then one block would consist of the 5 anxious students and about half (2 or 3) would be guaranteed to be assigned to the treatment. Figure 12.5 on the next page compares one possible randomization with no blocks to a blocking design.

► **FIGURE 12.5** (a) If there are no blocks, it is possible that all 5 of the higher-anxiety students (HA) will be assigned to the control group. (b) Creating blocks based on anxiety level ensures that roughly half of the higher-anxiety students (that is, 2 or 3 students) will be assigned to receive the treatment (T).

(a)	T	C	(b)	HA	LA
	LA	HA		T	T
	LA	LA		T	T
	LA	LA		T	T
	LA	HA		C	C
	LA	LA		T	T
	LA	LA		T	T
	LA	HA		C	C
	LA	HA		T	T
	LA	LA		T	T
	LA	LA		T	T
	LA	LA		T	T

Blocking and Matching

As we have just discussed, blocking combines units that are similar in order to improve the power of an experiment. What unit can be more similar than matching a person with himself or herself? Researchers can sometimes exploit the fact that measuring the same person twice (usually before and then after applying a treatment) results in positively correlated values. Studies that use this structure illustrate the **matched-pairs design**. In Chapter 9 you saw examples of such paired samples, and you learned how to apply the paired *t*-test to test hypotheses about differences between two groups of paired data. Pairs of data are, in fact, a form of blocking.

For example, in a study to detect the effects of the gastric band on weight loss (O'Brien 2010), patients had their weight measured without the band (before it was surgically put in place) and some time after the band was in place. In such a study, each patient is basically serving as his or her own control. The control group consists of the patients before the treatment. The treatment group consists of the *same* patients as in the control group, but after they've received the treatment. This sort of study design is typically more effective at measuring the effects of a treatment than one in which we have two groups of people (one group with no gastric band and the other with one) and compare the weights of the two groups.

EXAMPLE 4 Testing Diabetes Treatments

To combat diabetes, researchers compare two treatments designed to improve kidney functioning. One treatment uses both insulin and somatostatin (IS). The other uses insulin, somatostatin, and also glucagon (ISG). Each patient receives each treatment (with some number of days between treatments so that the effects of the first can wear off). After each treatment, the amount of urea nitrogen in the urine over a 24 hour period is measured. Low levels suggest the kidneys are not properly removing waste. The data are given in Table 12.1. Summary statistics are also provided. Assume that the subjects are randomly selected from a larger population and that the distribution of urea nitrogen levels in this population is approximately Normal.

► **TABLE 12.1** Outcome from diabetes study. IS stands for "insulin with somatostatin"; ISG stands for "insulin with somatostatin and also glucagon." Urea nitrogen amounts are in units of grams detected in urine over a 24 hour period. (Source: Rasking and Unger, 1978.)

Patient ID	IS (g/ 24 hr)	ISG (g/24 hr)	Difference
1	14	17	-3
2	6	8	-2
3	7	11	-4
4	6	9	-3
$\bar{x}$	8.25	11.25	-3.0
s	3.86	4.03	0.82

Looking Back

Paired Samples

Paired samples are those in which each object in one group is associated with one particular object in the other group. They are also sometimes called dependent samples or dependent observations.

QUESTIONS

- Perform a two-sample t -test (do not treat the data as paired data) to test whether the mean urea nitrogen levels are different under IS treatment and ISG. Assume the patients are a random sample from a larger population. Use a 5% significance level.
- Perform a paired t -test to test whether the mean urea nitrogen levels are different under IS treatment and under ISG treatment. Use a 5% significance level.
- Which approach is better? Why are the conclusions different?

SOLUTIONS

- We will quickly skip over the four steps of a hypothesis test so that we can emphasize the results of the two different procedures. Figure 12.6 shows the output from StatCrunch, which displays the null and alternative hypotheses, the observed value of the t -statistic, and the p -value.



Two sample T statistics with data

FIGURE 12.6 Output from two-sample t -test, no pairing.**Options****Hypothesis test results:** μ_1 : mean of var2 μ_2 : mean of var3 $\mu_1 - \mu_2$: mean difference $\mu_0 : \mu_1 - \mu_2 = 0$ $H_A : \mu_1 - \mu_2 \neq 0$

(without pooled variances)

Difference	Sample Mean	Std. Err.	Df	T-Stat	P-value
$\mu_1 - \mu_2$	-3	2.7913556	5.989039	-1.0747466	0.3239

From this we see that the value of the t -statistic is -1.07 and the p -value is 0.3239 . Because the p -value is bigger than 0.05 , we do not reject the null hypothesis. Our conclusion is that there is no evidence of a difference in mean urea nitrogen levels between the two treatments.

- Now we do a paired t -test. Figure 12.7 shows the StatCrunch output.



Paired T statistics

FIGURE 12.7 Output from paired t -test.**Options****Hypothesis test results:** $\mu_1 - \mu_2$: mean of the paired difference between var2 and var3 $\mu_0 : \mu_1 - \mu_2 = 0$ $H_A : \mu_1 - \mu_2 \neq 0$

Difference	Sample Diff.	Std. Err.	DF	T-Stat	P-value
$\mu_1 - \mu_2$	-3	0.4082483	3	-7.3484693	0.0052

This time the t -statistic is very different: -7.3 . And the p -value is much less than 0.05 . This time, we reject the null hypothesis and conclude that the treatments have different effects on the mean urea nitrogen levels.

- c. The data in this sample are paired; each patient contributes two data values to the sample. Because the values within pairs are positively correlated, the paired t -test has greater power than the two-sample t -test to detect differences in the population. The paired t -test is the more appropriate test for this situation. One clue to why the paired t -test has greater power in this situation can be seen by comparing the standard error for the paired test statistic with that for the two-sample statistic. The standard error for the paired statistic is much smaller, because with this study design, correctly using the paired test statistic results in a much more precise estimator.

Note that in both part a and part b, the differences in the sample means are the same: -3.0 . But our hypothesis is not about the sample, it is about the population. The paired t -test allowed us to see that the population means are different.

TRY THIS! Exercise 12.23



Another type of study that uses subjects as their own blocks is the **crossover design**. *Crossover design* is really just a more general term for the matched-pairs design. The difference is that sometimes there is more than one treatment, or more than one dosage of the same treatment. In a crossover design, researchers give all subjects the same set of treatments, but they usually randomly choose the order in which each subject receives the treatment.

Smoking causes many health problems, but quitting smoking can be difficult for some people. Researchers in New Zealand used a crossover design to compare three methods for helping people quit: snus (a Swedish chewing tobacco), nicotine gum, and Zonnic (which they described as an "oral nicotine sachet"). The goal was to deliver the nicotine that smokers craved in a form less harmful than cigarettes (Caldwell and Crane 2010). Patients used each treatment for two weeks. Patients were randomly assigned to the *order* in which they received each treatment. For example, some took snus first, then the gum, and then the Zonnic. Others also started with snus but then took Zonnic and ended with the gum. There were six different possible orderings to which subjects could be assigned.

During each two-week treatment, the patients recorded their desire for smoking. At the end of the trial, the researchers concluded that all three methods reduced the withdrawal symptoms created by quitting cigarettes, but that patients preferred Zonnic and snus over gum because there were fewer side effects.

SECTION 12.2

Controlling Variation in Surveys

In Chapter 7 we discussed a very basic method for collecting survey data called the simple random sample (SRS). In an SRS, we first create a list of all people in our population. Then we randomly choose people without replacement to participate in our study.

This basic method creates a sample that, if large enough, is representative of the population. Because it is based on a random scheme, we can compute a margin of error that measures how far away our estimate is likely to be from the population value, even when we don't know what the population value is.

In practice, however, simple random samples are difficult to carry out. In very large populations such as the United States (or any other country, for that matter), no single list of names exists for researchers to use. Also, this method can be imprecise: The margin of error can be big. In this section, we give an overview of other random sampling techniques that are more practical than simple random sampling.

Review of Sampling Basics

Before talking about more complex sampling schemes, let's take a moment to review some of the basics we introduced in Chapter 7.

One goal of a survey is to estimate the value of a population parameter on the basis of a sample. *Parameters* are numbers that characterize populations. The proportion of Americans who bought a new car in the last two years is a parameter, as is the mean weight of all eighth graders in your town. If we collect a representative sample from the population of interest, then a statistic from that sample can be used to estimate the parameter. A *statistic* is a number based on a sample of data. The average weight of 20 eighth graders chosen from your town is a statistic, as is the proportion of 100 people selected from the United States who say they bought a new car in the last two years.

Two features that measure the quality of an estimator are bias and precision. *Bias* measures how far away an estimator tends to be from the correct population value. If, on average, the estimator tends to get the value just right, then we say that the estimator is unbiased. *Precision* measures how the estimator varies from sample to sample. Different researchers using different random samples should get similar estimates when using a precise estimator. Precision is measured by the standard error, which is just the standard deviation the estimator would have if we were to repeat the survey many times. Ideally, an estimator is unbiased and has a small standard error.

Estimators can be biased for many different reasons. One common reason is that the sample is not representative of the population. For example, if we were to survey customers at a car dealership and ask whether they bought a new car in the last two years, these people would probably not be representative of the population as a whole. (They are probably less likely than the general population to have a new car at home if they're out shopping for a car.)

One way in which a nonrepresentative sample often arises is in situations where the sample chooses itself by volunteering to participate. For example, a political website advertised a poll on health care reform, and interested readers could click on the link to answer this question: "After reading President Obama's proposal, have your views about passing health care reform changed?" About 90 percent of 131,577 respondents said "No" (FoxNews.com/opinion, viewed February 25, 2010). Is this representative of the entire U.S. population? Most probably not. People chose whether to click on the link, and this decision is quite likely to be based on the fact that they have strong opinions on the topic, which they wish to share.

This situation is similar to an observational study. People choose whether or not to participate in the poll, so we don't know whether our estimate of the population parameter is representative only of the sort of person who would answer this poll or is representative of the population as a whole. (The fact that this health care poll appeared on a politically conservative website probably accounts in part for the large proportion who don't like the idea of passing health care reform. If the poll were on a liberal news site, the results might very well go in the opposite direction.)

The only way to collect data that avoids this potential bias is to select the sample randomly. A random sample has another advantage: It gives us a way to measure our confidence in the estimate. This measure takes the form of a confidence interval. A confidence interval enables us to say we're highly confident that the true population proportion is within, say, 3 percentage points of our estimate.

The precision of a survey estimator depends on the size of the sample and on the variability in the population. The good news is that the precision can always be improved by taking a larger sample (assuming you can afford to collect more data, that is). If everyone agrees and thinks the same way—in other words, if there is no variability—then the estimator will be extremely precise. The worst-case scenario occurs when the population is evenly split. If the true proportion in the population is $p = 0.50$, then the precision will be as bad as it gets.

For example, if $p = 0.50$ and we take a simple random sample of 100, then the standard error is

$$\sqrt{\frac{p(1-p)}{100}} = \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05.$$

A population with less diversity, say $p = 0.90$, would produce more precise estimators because the standard error is smaller:

$$\sqrt{\frac{p(1-p)}{100}} = \sqrt{\frac{0.9 \times 0.1}{100}} = 0.03.$$

Several approaches to collecting samples get around the difficulties posed by SRSs. These approaches either make it easier to access people to survey, or help by increasing the precision without the costs associated with a larger sample size.

Systematic Sampling

! Caution

Measurement Bias

Do not overlook the fact that estimates of the population proportion can also be biased if the questions are poorly worded.

Systematic sampling is one technique that can be used, in some situations, to obtain a random sample. In a systematic sample, objects from the population are sampled at regular intervals, such as taking every fifth person. One of the most common examples of systematic sampling occurs in exit polls held during elections.

The purpose of an exit poll is to estimate the winner of an election before the votes are counted. They are called exit polls because people are surveyed as they exit the polling place. The pollsters—the people collecting the data—are instructed to stop people at specific intervals. For example, they might stop every fifth person (or every tenth person) throughout the day.

Systematic sampling works well when you have subjects coming at you in some sort of sequence and when you have good reason to believe that the opinions will be randomly mixed during the time period when you collect that data. Exit polls would fail if, for example, pollsters collected data only first thing in the morning, or only in the middle of the afternoon. This is because it is possible that a certain type of voter goes to the polls before work, and if so, then a systematic sample taken only in the morning will be biased.

Systematic samples are used in quality control studies at manufacturing plants. Products come off an assembly line, and every tenth product might be inspected for defects. At the end of the day, this method yields a good estimate of the proportion of all products produced that day that have defects, as long as there is no particular pattern to the order in which defects arrive.

Stratified Sampling

Stratified sampling is a method for collecting data with increased precision. To perform stratified sampling, researchers first identify **strata**. Strata are similar to blocks; they are collections of people who are very similar to each other. Stratified sampling is based on a very simple principle: If we know that everyone in some group will answer the question the same way, then there's no need to ask everyone.

To create strata, choose a variable that identifies people who hold similar beliefs that are associated with the variable you are studying. For example, suppose you were designing a survey to determine the proportion of the population that supports allowing gay and lesbian citizens to serve in the military. According to Gallup Polls and other research, support for this idea depends on age: Younger people are more supportive than older people. Thus, even though we could take a large sample of 1000 people to get an approximately 3-point margin of error for a 95% confidence interval, we could also stratify on age and take a smaller sample to get the same precision.

A stratified sample would divide the population into strata based on age group. One stratum (the singular form of *strata*) might consist of 18–25-year-olds, another of

26–40-year-olds, and so on. We would then randomly select people (using a simple random sample) from each stratum and ask their opinion.

Stratified sampling works because it creates “mini-populations” that have lower variability than the population as a whole. If the strata were chosen carefully, then within each stratum, there is less variability than in the population as a whole. This means that within each stratum, we can produce a more precise estimator.

The result is that we have mini-estimates from each of the mini-populations. These estimates are each very precise, because the strata were chosen to have low variability within them. We then combine the mini-estimates into one estimate for the entire population. We can show mathematically that in many cases, this combined estimator is more precise.

KEY POINT

Stratified sampling improves the precision of estimators by sampling from within strata. Each stratum is created in such a way that people (or objects) within the stratum are similar to one another.

Age is one variable commonly used for creating strata. Political party (Republican or Democrat) is another. Other possibilities include gender, educational level, and geographic region. For example, if we were to take an opinion poll about the pop music star Justin Bieber, a performer with a large fan base of pre-teens, we could probably improve our survey by stratifying on age. One stratum might consist of people between the ages of 8 and 13, another of people aged 14–19, another of people aged 20–25, and so on. From each of these groups, we would select a random sample for our survey.

If the variable chosen to create the strata does not actually end up creating groups with similar-thinking members, then stratification will not offer any advantage over SRS. In fact, in some very unusual cases, it can make the precision worse! For this reason, it's important to choose the right stratification plan.

Cluster Sampling

Cluster sampling is a method that makes it easier to access very large populations. To perform a cluster sample, researchers divide the population into distinct groups, or **clusters**, using some sort of natural or convenient distinction. Next, they select a random sample (using SRS) of clusters and survey *everyone* within the clusters selected. For example, suppose you are interested in determining the number of children living in rental apartments in a city. Each apartment building can be a cluster, and you can then randomly select some number of apartment buildings and then survey everyone within those buildings to find out how many children are living in the apartments.

Clustering can happen over several stages. Randomly sampling, say, ten states in the United States and surveying everyone in these ten states is still a gigantic undertaking. Instead, after the states are selected, we might define a second stage of clusters consisting of cities, or maybe counties, zip codes, or the like. We would then take a random sample of these smaller clusters and survey everyone inside the clusters selected.

Foresters have used cluster sampling to estimate the extent to which bark beetles have infested forests. Bark beetles are a serious threat to forests, so being able to detect the proportion of damaged trees in areas in which there is not yet much damage is crucial to deterring their attack. By using satellite images, researchers can randomly select parcels of land to serve as clusters and then investigate every tree within those parcels (Coggins, Coops, and Wulder 2010). Cluster sampling is also used in humanitarian relief efforts to understand the extent to which people have been injured or killed by natural disasters. Researchers randomly select regions and then go door-to-door within them to collect damage reports.

Cluster sampling is sometimes confused with stratified sampling because both techniques break the population into groups. However, in some ways, cluster sampling

! Caution

Stratified Sampling and Blocking

Stratified sampling and blocking are similar. Both put subjects into groups (strata or blocks) so that subjects are similar within the group. However, stratified sampling is about ways of selecting subjects, and blocking is about a way of randomly assigning already selected subjects to treatment groups.

is the opposite of stratified sampling. In stratified sampling, the strata contain people who are all as similar as possible (for example, all Republicans in one strata and all Democrats in another). Cluster sampling, on the other hand, works best when each cluster by itself represents the entire population. This is rarely achieved, but when it is achieved, it ensures that estimates are as precise as possible.



Cluster sampling can be used to access very large populations. A cluster consists of a group of objects or people. To perform a cluster sample, we take a random sample of clusters and survey everyone within the selected clusters.

EXAMPLE 5 Counting Young Female Smokers

Health experts are concerned about the large number of smokers among young Chinese. However, experts do not know how many smokers there are between the ages of 14 and 24. To estimate the number of young, female smokers in China, researchers in California and China conducted a survey in which they randomly selected schools in China and then randomly selected classrooms in the selected schools. They surveyed all girls and young women in the selected classrooms and concluded that about 20% of all Chinese females in this age group were smokers (Ho et al. 2009).

QUESTION Is this a systematic sample, a stratified sample, or a cluster sample?

SOLUTION This is a cluster sample. More precisely, this is a two-stage cluster sample. The clusters are schools at one stage and then classrooms at the second stage.

TRY THIS! Exercise 12.29

SECTION 12.3

Reading Research Papers

One goal of this book is to teach you enough of the basic concepts of statistics that you can critically evaluate published research. The medical literature, in particular, records many findings that can have major consequences in our lives. When we rely on the popular media to interpret these findings, we often get contradictory messages. However, you now know enough statistics so that, with a few guiding principles, you can often make sense of research results yourself.

Before discussing ways of evaluating individual research articles, we offer a few over-arching, guiding principles.

1. *Pay attention to how randomness is used.* Random sampling is used to obtain a representative sample, so that we can make inferences about a larger population. Random assignment is used to test causal associations so that we can conclude that the treatment was truly effective (or was not) on a particular sample of subjects. Many medical studies use random assignment but do not use random sampling. This means the results are not necessarily applicable to the entire population. Table 12.2 summarizes these possibilities.
2. *Don't rely solely on the conclusions of any single paper.* Research advances in small steps. A single research study, even when the conclusions are grand and ambitious, can tell us only a small part of the real story. For example, the *Los Angeles Times* reported, on the basis of a few published studies, that many people considered vitamin D to be useful for preventing cancer, cardiovascular

	Sample Selected Randomly	No Random Sample
Random Assignment	You can make a causal conclusion and conclude that the entire population would be affected similarly.	You can make a causal conclusion, but we do not know whether everyone would respond similarly.
No Random Assignment	You can assume that an association between the variables exists in the population, but you cannot conclude that it is a causal relationship.	You can conclude that an association exists within the sample but not in the entire population, and you cannot conclude that there is a causal relationship.

TABLE 12.2 Four different study designs and the inferences possible.

disease, depression, and other maladies. But a panel of medical experts concluded that these beliefs were based on preliminary studies. The body of medical research, they concluded, was in fact quite mixed in its view of just how effective vitamin D really is (Healy 2011).

3. *Extraordinary claims require extraordinary evidence.* This advice was a favorite piece of wisdom of magician and professional skeptic (The Amazing) James Randi. If someone claims to have done something that had been believed impossible, don't believe it until you've seen some very compelling evidence.
4. *Be wary of conclusions based on very complex statistical or mathematical models.* Complex statistics often require complex assumptions, and one consequence of this is that the findings might be correct and true, but only in a limited set of circumstances. Also, some research studies are essentially "what if" studies: "We're studying *what* will happen *if* these assumptions hold." It is important to remember that those assumptions may not be practical—or even possible.
5. *Stick to peer-reviewed journals.* **Peer review** means that papers were read by two or three (and sometimes more) knowledgeable and experienced researchers in the same field. These reviewers can prevent papers from being published if they do not think the methods employed were sound or if they find too many mistakes. They can also demand that the authors make changes and submit the paper again. Many published papers have gone through several rounds of reviews. But be careful: Not all peer-reviewed journals are equal. The more prestigious journals have greater resources to check papers for mistakes, and they have much more careful and knowledgeable reviewers on hand to find sometimes subtle errors. Good journals have editorial boards that reflect the general practice of the community and are not restricted to people who reflect a minority point of view.

EXAMPLE 6 Improving Tips

A sociologist wonders whether a waitress's tip is affected by whether or not she touches her customers. Below are two study designs that the sociologist might use. Read these, and then answer the questions that follow.

Design A: The sociologist chooses four large restaurants in his city and gets all waitresses at these restaurants to agree to participate in his study. On several nights during the next few weeks, the sociologist visits the restaurants and records the number of times all waitresses touch their customers on the customer's back or shoulder. He also records the total amount of tips earned by each waitress. He finds that waitresses who sometimes touched their customers earned larger tips, on average, than those who did not.

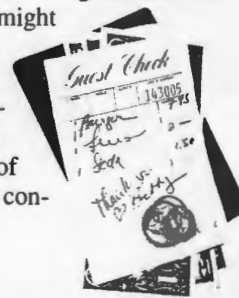
Design B: The sociologist chooses four large restaurants in his city and gets all waitresses at these restaurants to agree to participate in his study. At each restaurant, half of the waitresses are randomly assigned to either the “touch” group or the “no touch” group. The “touch” group is instructed to lightly touch each customer on the back or shoulder two or three times during the meal. The “no touch” group is instructed not to touch customers at all. At the end of the study, the sociologist finds that waitresses in the “touch” group earned larger tips, on average, than those in the “no touch” group.

QUESTION For each study design, state whether it is possible to generalize the results to a larger population and whether the researcher can make a cause-and-effect conclusion.

SOLUTION For Design A, the waitresses (or restaurants) were not randomly selected from a larger population, so we cannot generalize the results beyond the sample. This was an observational study: The waitresses themselves chose whether to be “touchers” or “no touchers.” Because random assignment was not used, we cannot conclude that the difference in tips was caused by touching. A possible confounding effect is the restaurant itself, which might encourage touching and which might have a clientele that tends to tip more generously than the clientele at other restaurants.

For Design B, we also cannot generalize to a larger population. But here random assignment was used. The researcher blocks on restaurant, so he also controls for the environment of the restaurant. Because random assignment was used, we can conclude that the touching caused the increased amount of tips.

TRY THIS! Exercise 12.35



Reading Abstracts

An **abstract** is a short paragraph at the beginning of a research article that describes the basic findings. If you look up science papers using Google Scholar, say, clicking on the link will usually take you to an abstract. Often, you will be able to read the abstract at no charge. (Reading the papers themselves often requires that you subscribe to the journal or that you read from the computers at a library that has purchased a subscription.) In addition, the websites of many journals display the abstracts to their published papers, even if they do not provide access to the papers themselves.

For example, an article recently published in the *New England Journal of Medicine* (Poordad, et al., 2011) reported the findings of a study on the effectiveness of a new treatment for chronic hepatitis C virus (HCV). Currently, the standard treatment, peginterferon-ribavirin, has a relatively low success rate. A success, in treating HCV, is called a “sustained virologic response.” In a sustained virologic response, the virus is not eliminated from the body, but is undetectable for a long period of time. Researchers hope that adding a new medicine, boceprevir, to the standard treatment would lead to a great proportion of patients achieving a sustained virologic response. An excerpt from the abstract follows:

Methods: We conducted a double-blind study in which previously untreated adults with HCV genotype 1 infection were randomly assigned to one of three groups. In all three groups, peginterferon alfa-2b and ribavirin were administered for 4 weeks (the lead-in period). Subsequently, group 1 (the control group) received a placebo plus peginterferon-ribavirin for 44 weeks; group 2 received boceprevir plus peginterferon-ribavirin for 24 weeks, and those with a detectable HCV RNA level between weeks 8 and 24 received a placebo plus peginterferon-ribavirin for an additional 20 weeks; and group 3 received boceprevir plus

peginterferon-ribavirin for 44 weeks. Nonblack patients and black patients were enrolled and analyzed separately.

Results: A total of 938 nonblack and 159 black patients were treated. In the nonblack cohort, a sustained virologic response was achieved in 125 of the 311 patients (40%) in group 1, in 211 of the 316 patients (67%) in group 2 ($P < 0.001$), and in 213 of the 311 patients (68%) in group 3 ($P < 0.001$). In the black cohort, a sustained virologic response was achieved in 12 of the 52 patients (23%) in group 1, in 22 of the 52 patients (42%) in group 2 ($P = 0.04$), and in 29 of the 55 patients (53%) in group 3 ($P = 0.004$). In group 2, a total of 44% of patients received peginterferon-ribavirin for 28 weeks. Anemia led to dose reductions in 13% of controls and 21% of boceprevir recipients, with discontinuations in 1% and 2%, respectively.

Conclusions: The addition of boceprevir to standard therapy with peginterferon-ribavirin, as compared with standard therapy alone, significantly increased the rates of sustained virologic response in previously untreated adults with chronic HCV genotype 1 infection. The rates were similar with 24 weeks and 44 weeks of boceprevir. (Funded by Schering-Plough [now Merck]; SPRINT-2 ClinicalTrials.gov, number NCT00705432.)

To help you evaluate this abstract, and others like it, answer these questions:

1. What is the research question that these investigators are trying to answer?
2. What is their answer to the research question?
3. What were the methods they used to collect data?
4. Is the conclusion appropriate for the methods used to collect data?
5. To what population do the conclusions apply?
6. Have the results been replicated (that is, reproduced) in other articles? Are the results consistent with what other researchers have suggested?

The sixth question is important, although it is often difficult for a layperson to answer. Research is a difficult activity, in part because of the great variability in nature. Studies are often done with samples that do not allow generalizing to populations or are subject to bias, and sometimes researchers just make mistakes. For that reason, you should not believe in the claims of a single study unless it is consistent with currently accepted theory and supported by other research. For all of these reasons, wait until there is some accumulation of knowledge before changing your lifestyle or making a major decision on the basis of scientific research.

Let's see how we would answer these questions for this abstract.

1. *What is the research question that these investigators are trying to answer?* Does the addition of boceprevir to the standard treatment lead to a higher proportion of Hepatitis C patients achieving a sustained virologic response?
2. *What is their answer to the research question?* Yes, the additional drug improved patients' responses. Researchers report an increased rate of patients achieving sustained virologic response with boceprevir plus the standard therapy compared to those receiving only standard therapy.
3. *What were the methods they used to collect data?* Patients were randomly assigned to one of three treatment groups. Assignments were double-blind. Patients were examined after 24 and 44 weeks to determine if they had achieved sustained virologic response.
4. *Is the conclusion appropriate for the methods used to collect data?* The fact that subjects were randomly assigned to treatment groups means that we can make a causal conclusion and say that the difference in virologic response rates was due to the addition of boceprevir to the standard

treatment. We also know that the observed results were too large to be explained by chance. Note that p-values are given as, for example, " $P < 0.001$ ". For instance, we know that group 2 (nonblack cohort) had a greater rate of sustained virologic response than the treatment group, with $p\text{-value} < 0.001$.

5. *To what population do the conclusions apply?* The sample was not randomly selected, so although we *can* conclude that the treatment was effective, we do not know if we would see the same sized effect on other samples or in the population of all Hepatitis C patients. Because it is known that African Americans and non-African Americans have different responses to the standard treatment, these two groups were treated separately in the analysis.
6. *Have the results been replicated in other articles? Are the results consistent with what other researchers have suggested?* We cannot tell from the abstract.

Note that even though we don't know which statistical procedures were carried out (we could learn this from reading the article itself), we still have a good sense of the reliability of the study. We know that the reported success rates for standard treatment for Hepatitis C (40% in the nonblack cohort and 23% in the black cohort) were lower than for the treatment that included boceprevir (67% for nonblack cohort after 24 weeks, 68% after three weeks; 42% for the black cohort after 24 weeks, and 53% after 44 weeks.) We know this difference cannot be plausibly assigned to chance, and we know that, because of random assignment, that boceprevir was the cause of the difference in rates.

EXAMPLE 7 Driving While Talking on a Cell Phone

Read the abstract that follows, which concerns the hazards of driving while talking on a cell phone (Strayer et al. 2006).

Objective: The objective of this research was to determine the relative impairment associated with conversing on a cellular telephone while driving. **Background:** Epidemiological evidence suggests that the relative risk of being in a traffic accident while using a cell phone is similar to the hazard associated with driving with a blood alcohol level at the legal limit. The purpose of this research was to provide a direct comparison of the driving performance of a cell phone driver and a drunk driver in a controlled laboratory setting. **Method:** We used a high-fidelity driving simulator to compare the performance of cell phone drivers with drivers who were intoxicated from ethanol (i.e., blood alcohol concentration at 0.08% weight/volume). **Results:** When drivers were conversing on either a hand-held or hands-free cell phone, their braking reactions were delayed and they were involved in more traffic accidents than when they were not conversing on a cell phone. By contrast, when drivers were intoxicated from ethanol, they exhibited a more aggressive driving style, following closer to the vehicle immediately in front of them and applying more force while braking. **Conclusion:** When driving conditions and time on task were controlled for, the impairments associated with using a cell phone while driving can be as profound as those associated with driving while drunk. **Application:** This research may help to provide guidance for regulation addressing driver distraction caused by cell phone conversations.

QUESTION Write a short paragraph describing this research. Use the six questions above for guidance.

SOLUTION The study asks whether talking on a cell phone impairs driving and, if so, whether the impairment is similar to the impairment caused by driving drunk. The researchers conclude that the impairment is as great when talking on a cell phone as when driving drunk.

drivers use a driving simulator. Some drivers talked on their phone during the task, and others were given sufficient alcohol to make them legally drunk. However, these techniques as described in the abstract do not lead to much faith in the conclusion.

- We do not know whether drivers were randomly assigned to treatment groups.
- We do not know how these drivers were chosen.
- We do not know how many drivers were in each group.
- Although we know that all of the drivers in the alcohol group had the same blood alcohol level, we do not know whether all of the drivers in the cell phone group were equally engaged in the task. For example, perhaps some had conversations that required more concentration.
- We also do not know whether the results are due to chance, because no statistical conclusions are provided.
- Presumably, this finding is meant to apply to all cell phone drivers, but we are not given enough information about the subjects to know how they were selected, so we don't know how well we can generalize these results to a larger population.

Despite these shortcomings, the abstract states that the study's findings are consistent with a larger body of literature that finds that cell phone driving is dangerously distracting.



TRY THIS Exercise 12.45

Buyer Beware

Tolstoy's novel *Anna Karenina* begins, "Happy families are all alike; every unhappy family is unhappy in its own way." To (very loosely) paraphrase the great Russian novelist, there are few ways that a study can be good, but there are many ways that studies can go wrong. We've given you some tips that should help you recognize a good study, but you should also be aware of some warning signs and features that indicate that a study might not be good.

Data Dredging

Hypothesis testing is designed to test claims that result from a theory. The theory makes predictions about what we should see in the data; for example, students who write about their anxieties will do better on an exam, so the mean exam score of students who wrote about their anxieties should be greater than the mean score of those who did not. We next collect the data and then do a hypothesis test to determine whether the theory was correct. **Data dredging** is the practice of stating our hypotheses after first looking at the data. Data dredging makes it more likely that we will mistakenly reject the null hypothesis. Even when the null hypothesis is true, our data sometimes show surprising outcomes, just because of chance variation. If we first look at the test statistic to decide what the hypotheses should be, we are rigging the system in favor of the alternative hypothesis.

The situation is analogous to betting on a horse race after the race has begun. You are supposed to place a bet before the race starts, so that everyone is on equal footing. The odds on which horse will win (which determine the payoffs) are meant to estimate probabilities of a horse winning. But if you wait until the horses are running, you have a better chance than everyone else of correctly predicting which horse will win. You have "snooped" at the data to make your decision. Your probability of winning is not the same as what everyone else believed it to be.

Theories should be based on data, but the correct procedure is to use data to formulate a theory and then to collect additional data in an independent study to test that theory. If it is too costly to collect additional data, one common approach is to

randomly split the data set into two (or even more) pieces. One piece of the data is then set aside—"locked away"—and the researchers are forbidden to look. The researchers can then examine the first piece as much as they want and use these data to generate hypotheses. After they have generated hypotheses, they can test these hypotheses on the second, "locked away" data set. Another possibility is to use the data for generating hypotheses and then go out and collect *more* data.

Publication Bias

Most scientific and medical journals prefer to publish "positive" findings. A positive finding is one in which the null hypothesis is rejected (with the result that the researcher concludes that the tested treatment is effective). Some journals prefer this sort of finding, because these are generally the results that advance science. However, suppose a pharmaceutical company produces a new drug that it claims can cure depression, and suppose this drug does not work. If many researchers are interested in studying this drug and they do statistical tests with a 5% significance level, then 5% of them will conclude that the drug is effective, even though we know it is not. If a journal favors positive findings over negative findings, then we will read only about the studies that find the drug works, even though the vast majority of researchers came to the opposite conclusion.

Publication bias is one reason why it is important to consider several different studies of the same drug before making decisions. A new, and somewhat controversial, form of statistical analysis called meta-analysis has been developed in recent years, in part to help with problems such as these. A **meta-analysis** considers all studies done to test a particular treatment and tries to reconcile different conclusions, attempting to determine whether other factors, such as publication bias, played a role in the reported outcomes.

Psychologists conducted a meta-analysis to conclude that violent video games are not associated with violent behavior in children, despite the fact that several studies had concluded otherwise (Ferguson and Kilburn 2010). They concluded that papers that found an association between video game violence and real-life violence were more likely to be published, so researchers who found no such associations were less likely to be published. This result is not likely to settle the controversy, but it points to a potential danger that arises when journals publish only positive findings.

Profit Motive

Much statistical research is now paid for by corporations that hope to establish that their products make life better for people. Researchers are usually required to disclose to a journal whether they are themselves making money off the drug or product that they are researching, but this does not tell us who funded the project in the first place. There is no reason to reject the conclusions of a study simply because it was paid for by a corporation or business or some other organization with a vested interest. You should always evaluate the methods of the study used and decide whether the methods are sound.

However, you should be aware that sometimes the corporation funding the research can influence whether results get published and which results get published. For example, a researcher might be funded by a pharmaceutical company to test a new drug. If he finds the drug doesn't work, the drug company might decide that it doesn't want to publicize this fact. Or perhaps the drug works on only a small subset of people. The company might then publicize that the drug is effective but fail to mention that it is effective on only a small group.

For example, a 2007 study concluded that playing "active" video games (such as the Wii) was healthier than playing "passive" games and found that children burned more calories playing the active games (Neale 2007). The study also suggested that playing real sports was much more healthful than playing any video game, but still the researchers conclude, "Nevertheless, new generation computer games stimulated

obesity, such positive behaviors should be encouraged.” A press release notes that the study was funded by Cake, the “marketing arm” of Nintendo UK, which manufactures the Wii. This does not mean that the results of the study are wrong (the children playing the active games most likely did in fact burn more calories than those who played the passive games), but it may account for the positive spin despite the finding that active video games were not nearly as effective as playing sports at promoting weight loss.

Media

The media—newspapers, magazines, television shows, and radio broadcasts—are also profit-motivated. A good journalist strives to get to the heart of the matter. Nonetheless, scientific and medical research findings are complex, and when condensing complex ideas into easily digested sound bites (and doing so in a way that entertains those who pay for the papers and magazines), the truth sometimes gets distorted.

The media often use catchy headlines, and these headlines do not always capture the true spirit of a study. The most common problem is that headlines often suggest a cause-and-effect relationship, even though the wise statistics student will quickly recognize that such a conclusion is not supported by the data. Statistician Jonathan Mueller keeps a website of such headlines. A couple of examples: “Studies say lots of candy could lead to violence.” “Texting lifts crash risk by large margin.” These headlines all suggest causality: Eating candy will make you violent. Texting will increase the risk of a car accident. But you now know that such a conclusion can be made only for controlled experiments, and even then we must be cautious. Ideally, you can learn what you need to know by reading the news article. But not always. The information you need to judge whether the conclusions of a study are strong is often missing from news reports.

Clinical Significance vs. Statistical Significance

An outcome of an experiment or study that is large enough to have a real effect on people’s health or lifestyle is said to have **clinical significance**. Sometimes, researchers discover that a treatment is statistically significant (meaning that the outcome is too large to be due to chance) but too small to be meaningful (so it is not clinically significant). Studies with very large sample size have large power and so are capable of detecting even very small differences between treatment groups. This does not mean the differences matter. For example, a drug might truly lower cholesterol levels, but not enough to make a real difference in someone’s health. Or playing Wii might burn more calories, but maybe not enough for it to serve as a form of exercise. Treatments that cause meaningful effects are called clinically significant. Sometimes, statistically significant results are not clinically significant.

Imagine a rare disease that only 1 person in 10 million people gets. A controlled experiment finds that a new drug “significantly” reduces your risk of getting this disease; specifically, it cuts the risk in half. Here, *significantly* means that the reduction in risk is statistically significant. But given that the disease is so rare already, is it worth taking medicine to cut your risk from 1 in 10 million to 1 in 20 million? Particularly if the drug is expensive or has side effects, most people would probably decide that this treatment is not clinically significant.

CASE STUDY REVISITED

At the beginning of this chapter, we reported on a study that found that stretching before a race did not increase a runner’s speed. Here is the abstract:

The results of previous research have shown that passive muscle stretching can diminish the peak force output of subsequent maximal isometric, concentric and stretch-shortening contractions. The aim of this study was to establish whether the deleterious effects of passive stretching seen in laboratory settings would be manifest in a performance setting. Sixteen members (11 males, 5 females) of a Division I NCAA