# Experiments, Good and Bad

5

© Jose Luis Pelaez, Inc./CORBIS

**CASE STUDY** Reports about climate change appear frequently in the media. Climate scientists warn us that major changes will occur in the coming years. For example, scientists predict that the changing climate will probably bring more rain to California, but they don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Is it possible to investigate the effects of possible future changes in climate now?

Researchers at the University of California, Berkeley carried out an experiment to study the effects of more rain in either season. They randomly assigned plots of open grassland to three treatments. One treatment was to add water equal to 20% of annual rainfall during January to March (winter). A second treatment was to add water equal to 20% of annual rainfall during April to June (spring). The third treatment was to add no water beyond normal rainfall. Eighteen circular plots of area 70 square meters were used for this study, with six plots used for each treatment. One variable the researchers measured was total plant biomass, in grams per square meter, produced in a plot over a year. Total plant biomass for the three treatments was compared to assess the effect of increased rainfall.

Is this a good study? By the end of this chapter, you will be able to determine the strengths and weaknesses of a study such as this.



## Talking about experiments

Observational studies passively collect data. We observe, record, or measure, but we don't interfere. Experiments actively produce. Experimenters intentionally intervene by imposing some treatment in order to see what happens. All experiments and many observational studies are interested

in the effect that one variable has on another variable. Here is the vocabulary we use to distinguish the variable that acts from the variable that is acted upon.

### Vocabulary

A **response variable** is a variable that measures an outcome or result of a study.

An **explanatory variable** is a variable that we think explains or causes changes in the response variable.

The individuals studied in an experiment are often called **subjects**.

A **treatment** is any specific experimental condition applied to the subjects. If an experiment has several explanatory variables, a treatment is a combination of specific values of these variables.

### EXAMPLE 1 Learning on the Web

An optimistic account of learning online reports a study at Nova Southeastern University, Fort Lauderdale, Florida. The authors of the study claim that students taking undergraduate courses online were "equal in learning" to students taking the same courses in class. Replacing college classes with websites saves colleges money, so this study seems to suggest we should all move online.

College students are the *subjects* in this study. The *explanatory variable* considered in the study is the setting for learning (in class or online). The *response variable* is a student's score on a test at the end of the course. Other variables were also measured in the study, including the score on a test on the course material before the courses started. Although this was not used as an explanatory variable in the study, prior knowledge of the course material might affect the response, and the authors wished to make sure this was not the case.

### EXAMPLE 2 The effects of a sexual assault resistance program

Young women attending universities may be at risk of being sexually assaulted, primarily by male acquaintances. In an attempt to develop an effective strategy to reduce this risk, three universities in Canada investigated the effectiveness of a sexual assault resistance program. The program consists of four 3-hour units in which information is


provided and skills are taught and practiced, with the goal of being able to assess risk from acquaintances, overcome emotional barriers in acknowledging danger, and engage in effective verbal and physical self-defense.

First-year female students were randomly assigned to the program or to a session providing access to brochures on sexual assault (as was common university practice). The result was that the sexual assault resistance program significantly reduced rapes as reported during one year of follow-up.

The Canadian study is an experiment in which the *subjects* are the 893 first-year female students. The experiment compares two *treatments*. The *explanatory variable* is the treatment a student received. Several *response variables* were measured. The primary one was rape as reported by participants after a one-year follow-up period.

You will often see explanatory variables called *independent variables* and response variables called *dependent variables*. The idea is that the response variables depend on the explanatory variables. We avoid using these older terms, partly because "independent" has other and very different meanings in statistics.

## How to experiment badly

Do students who take a course via the Web learn as well as those who take the same course in a traditional classroom? The best way to find out is to assign some students to the classroom and others to the Web. That's an experiment. The Nova Southeastern University study was not an experiment because it imposed no treatment on the student subjects. Students chose for themselves whether to enroll in a classroom or an online version of a course. The study simply measured their learning. It turns out that the students who chose the online course were very different from the classroom students. For example, their average score on a test on the course material given before the courses started was 40.70, against only 27.64 for the classroom students. It's hard to compare in-class versus online learning when the online students have a big head start. The effect of online versus in-class instruction is hopelessly mixed up with influences lurking in the background. Figure 5.1 shows the mixed-up influences in picture form.

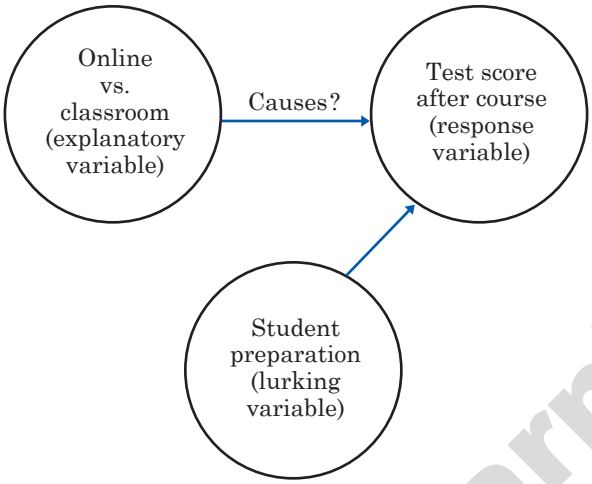



**Figure 5.1** Confounding in the Nova Southeastern University study. The influence of course setting (the explanatory variable) cannot be distinguished from the influence of student preparation (a lurking variable).

### Lurking variables

A **lurking variable** is a variable that has an important effect on the relationship among the variables in a study but is not one of the explanatory variables studied.

Two variables are **confounded** when their effects on a response variable cannot be distinguished from each other. The confounded variables may be either explanatory variables or lurking variables.

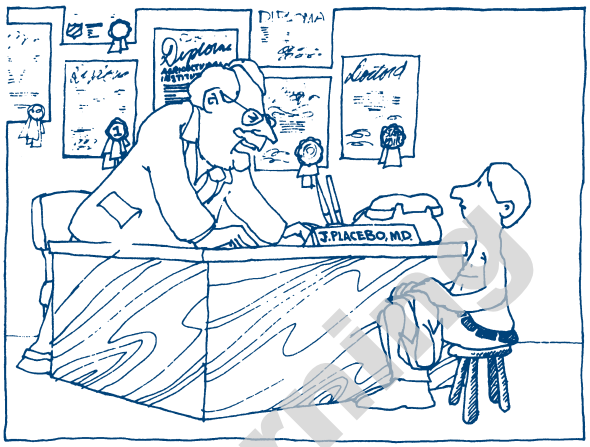
In the Nova Southeastern University study, student preparation (a lurking variable) is confounded with the explanatory variable. The study report claims that the two groups did equally well on the final test. We can't say how much of the online group's performance is due to their head start. That a group that started with a big advantage did no better than the more poorly prepared classroom students is not very impressive evidence of the wonders of Web-based instruction. Here is another example, one in which a second experiment was proposed to untangle the confounding.

### EXAMPLE 3 Pig whipworms and the need for further study

Crohn's disease is a chronic inflammatory bowel disease. An experiment reported in *Gut*, a British medical journal, claimed that a drink

containing thousands of pig whipworm eggs was effective in reducing abdominal pain, bleeding, and diarrhea associated with the disease.

Experiments that study the effectiveness of medical treatments on actual patients are called **clinical trials**. The clinical trial that suggested that a drink made from pig whipworm eggs might be effective in relieving the symptoms of Crohn's disease had a "one-track" design—that is, only a single treatment was applied:




"I want to make one thing perfectly clear, Mr. Smith. The medication I prescribe *will* cure that run-down feeling."

**Impose treatment ⟶ Measure response**

**Pig whipworms ⟶ Reduced symptoms?**

The patients did report reduced symptoms, but we can't say that the pig whipworm treatment caused the reduced symptoms. It might be just the **placebo effect**. A **placebo** is a dummy treatment with no active ingredients. Many patients respond favorably to *any* treatment, even a placebo. This response to a dummy treatment is the placebo effect. Perhaps the placebo effect is in our minds, based on trust in the doctor and expectations of a cure. Perhaps it is just a name for the fact that many patients improve for no visible reason. The one-track design of the experiment meant that the placebo effect was confounded with any effect the pig whipworm drink might have.

The researchers recognized this and urged further study with a better-designed experiment. Such an experiment might involve dividing subjects with Crohn's disease into two groups. One group would be treated with the pig whipworm drink as before. The other would receive a placebo. Subjects in both groups would not know which treatment they were receiving. Nor would the physicians recording the symptoms of the subjects know which treatment a subject received so that their diagnosis would not be influenced by such knowledge. An experiment in which neither subjects nor physicians recording the symptons know which treatment was received is called **double-blind**.

Both observational studies and one-track experiments often yield useless data because of confounding with lurking variables. It is hard to avoid confounding when only observation is possible. Experiments offer

better possibilities, as the pig whipworm experiment shows. This experiment could be designed to include a group of subjects who receive only a placebo. This would allow us to see whether the treatment being tested does better than a placebo and so has more than the placebo effect going for it. Effective medical treatments pass the "placebo test" by outperforming a placebo.

## Randomized comparative experiments

The first goal in designing an experiment is to ensure that it will show us the effect of the explanatory variables on the response variables. Confounding often prevents one-track experiments from doing this. The remedy is to *compare* two or more treatments. When confounding variables affect all subjects equally, any systematic differences in the responses of subjects receiving different treatments can be attributed to the treatments rather than to the confounding variables. This is the idea behind the use of a placebo. All subjects are exposed to the placebo effect because all receive some treatment. Here is an example of a new medical treatment that passes the placebo test in a direct comparison.

### EXAMPLE 4 Sickle-cell anemia

Sickle-cell anemia is an inherited disorder of the red blood cells that in the United States affects mostly blacks. It can cause severe pain and many complications. The National Institutes of Health carried out a clinical trial of the drug hydroxyurea for treatment of sickle-cell anemia. The subjects were 299 adult patients who had had at least three episodes of pain from sickle-cell anemia in the previous year. An episode of pain was defined to be a visit to a medical facility that lasted more than four hours for acute sickling-related pain. The measurement of the length of the visit included all time spent after registration at the medical facility, including the time spent waiting to see a physician.

Simply giving hydroxyurea to all 299 subjects would confound the effect of the medication with the placebo effect and other lurking variables such as the effect of knowing that you are a subject in an experiment. Instead, approximately half of the subjects received hydroxyurea, and the other half received a placebo that looked and tasted the same. All subjects were treated exactly the same (same schedule of medical checkups, for example) except for the content of the medicine they took. Lurking variables, therefore, affected both groups equally and should not have caused any differences between their average responses.

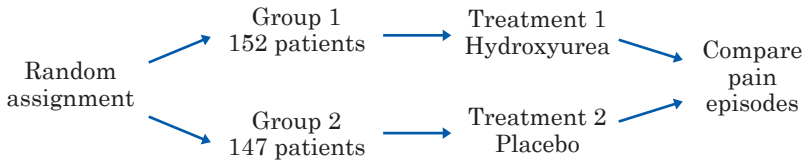



**Figure 5.2** The design of a randomized comparative experiment to compare hydroxyurea with a placebo for treating sickle-cell anemia, Example 4.

The two groups of subjects must be similar in all respects before they start taking the medication. Just as in sampling, the best way to avoid bias in choosing which subjects get hydroxyurea is to allow impersonal chance to make the choice. A simple random sample of 152 of the subjects formed the hydroxyurea group; the remaining 147 subjects made up the placebo group. Figure 5.2 outlines the experimental design.

The experiment was stopped ahead of schedule because the hydroxyurea group had many fewer pain episodes than the placebo group. This was compelling evidence that hydroxyurea is an effective treatment for sickle-cell anemia, good news for those who suffer from this serious illness.

Figure 5.2 illustrates the simplest **randomized comparative experiment**, one that compares just two treatments. The diagram outlines the essential information about the design: random assignment to groups, one group for each treatment, the number of subjects in each group (it is generally best to keep the groups similar in size), what treatment each group gets, and the response variable we compare. Random assignment of subjects to groups can use any of the techniques discussed in Chapter 2. For example, we could choose a simple random sample, labeling the 299 subjects 1 to 299, then using software to select the 152 subjects for Group 1. The remaining 147 subjects form Group 2. Lacking software, label the 299 subjects 001 to 299 and read three-digit groups from the table of random digits (Table A) until you have chosen the 152 subjects for Group 1. The remaining 147 subjects form Group 2.

The placebo group in Example 4 is called a **control group** because comparing the treatment and control groups allows us to control the effects of lurking variables. A control group need not receive a dummy treatment such as a placebo. In Example 2, the students who were randomly assigned to the session providing access to brochures on sexual assault (as was common university practice) were considered to be a control group. Clinical trials often compare a new treatment for a medical condition—not with a

placebo, but with a treatment that is already on the market. Patients who are randomly assigned to the existing treatment form the control group. To compare more than two treatments, we can randomly assign the available experimental subjects to as many groups as there are treatments. Here is an example with three groups.

### EXAMPLE 5 Conserving energy

Many utility companies have introduced programs to encourage energy conservation among their customers. An electric company considers placing electronic meters in households to show what the cost would be if the electricity use at that moment continued for a month. Will meters reduce electricity use? Would cheaper methods work almost as well? The company decides to design an experiment.

One cheaper approach is to give customers an app and information about using the app to monitor their electricity use. The experiment compares these two approaches (meter, app) and also a control. The control group of customers receives information about energy conservation but no help in monitoring electricity use. The response variable is total electricity used in a year. The company finds 60 single-family residences in the same city willing to participate, so it assigns 20 residences at random to each of the three treatments. Figure 5.3 outlines the design.

To carry out the random assignment, label the 60 households 1 to 60; then use software to select an SRS of 20 to receive the meters. From those not selected, use software to select the 20 to receive the app. The remaining 20 form the control group. Lacking software, label the 60 households 01 to 60. Enter Table A to select an SRS of 20 to receive the meters. Continue in Table A, selecting 20 more to receive the app. The remaining 20 form the control group.

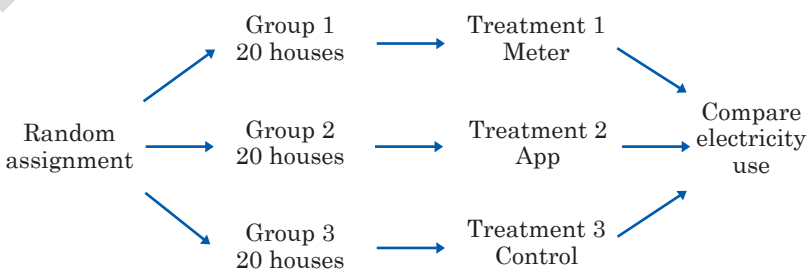



**Figure 5.3** The design of a randomized comparative experiment to compare three programs to reduce electricity use by households, Example 5.

**NOW IT'S YOUR TURN**

**5.1 Improving students' long-term knowledge retention.** Can a technique known as personalized review improve students' ability to remember material learned in a course? To answer this question, a researcher recruited 200 eighth-grade students who were taking introductory Spanish. She assigned 100 students to a class training students to use the personalized review technique as part of their study habits. The other 100 continued to use their usual study habits. The researcher examined both groups on Spanish vocabulary four weeks after the school year ended. Outline the design of this study using a diagram like Figures 5.2 and 5.3.

## The logic of experimental design

The randomized comparative experiment is one of the most important ideas in statistics. It is designed to allow us to draw cause-and-effect conclusions. Be sure you understand the logic:

- Randomization produces groups of subjects that should be similar, on average, in all respects before we apply the treatments.
- Comparative design exposes all groups to similar conditions, other than the treatments they receive. This ensures that any additional lurking variables operate equally on all groups and, on average, that groups differ only in the treatments they receive.
- Therefore, differences in the response variable must be due to the effects of the treatments.

We use chance to choose the groups in order to eliminate any systematic bias in assigning the subjects to groups. In the sickle-cell study, for example, a doctor might subconsciously assign the most seriously ill patients to the hydroxyurea group, hoping that the untested drug will help them. That would bias the experiment against hydroxyurea. Choosing an SRS of the subjects to be Group 1 gives everyone the same chance to be in either group. We expect the two groups to be similar in all respects—age, seriousness of illness, smoker or not, and so on. Chance tends to assign equal numbers of smokers to both groups, for example, even if we don't know which subjects are smokers.

What about the effects of lurking variables not addressed by randomization—for example, those that arise after subjects have been randomly assigned to groups? The placebo effect is such a lurking variable. Its effect occurs only after the treatments are administered to subjects. If the groups are treated at different times of the year, so that some groups are treated during flu season and others not, higher exposure of some groups to the flu

could be a lurking variable. In a comparative design, we try to ensure that these lurking variables operate similarly on all groups. All groups receive some treatment in order to ensure they are equally exposed to the placebo effect. All groups receive treatment at the same time, so all experience the same exposure to the flu.

It may not surprise you to learn that medical researchers adopted randomized comparative experiments only slowly—many doctors think they can tell "just by watching" whether a new therapy helps their patients. Not so. There are many examples of medical treatments that became popular on the basis of one-track experiments and were shown to be worth no more than a placebo when some skeptic tried a randomized comparative experiment. One search of the medical literature looked for therapies studied both by proper comparative trials and by trials with "historical controls." A study with historical controls compares the results of a new treatment, not with a control group, but with how well similar patients have done in the past. Of the 56 therapies studied, 44 came out winners with respect to historical controls. But only 10 passed the placebo test in proper randomized comparative experiments. Expert judgment is too optimistic even when aided by comparison with past patients. At present, U.S. law requires that new drugs be shown to be both safe and effective by randomized comparative trials. There is no such requirement for other medical treatments, such as surgery. A Web search of "comparisons with historical controls" found recent studies for other medical treatments that have used historical controls.

There is one important caution about randomized experiments. Like random samples, they are subject to the laws of chance. Just as an SRS of voters might, by bad luck, choose people nearly all of whom have the same political party preference, a random assignment of subjects might, by bad luck, put nearly all the smokers in one group. We know that if we choose *large* random samples, it is very likely that the sample will match the population well. In the same way, if we use *many* experimental subjects, it is very likely that random assignment will produce groups that match closely. More subjects means that there is less chance variation among the treatment groups and less chance variation in the outcomes of the experiment. "Use enough subjects" joins "compare two or more treatments" and "randomize" as a basic principle of statistical design of experiments.



### Principles of experimental design

The basic principles of statistical design of experiments are:

**1. Control** the effects of lurking variables on the response by ensuring all subjects are affected similarly by these lurking variables. Then simply compare two or more treatments.

**2. Randomize**—use impersonal chance to assign subjects to treatments so treatment groups are similar, on average.

**3. Use enough subjects** in each group to reduce chance variation in the results.

## Statistical significance

The presence of chance variation requires us to look more closely at the logic of randomized comparative experiments. We cannot say that *any* difference in the average number of pain episodes between the hydroxyurea group and the control group must be due to the effect of the drug. Even if both treatments are the same, there will always be some chance differences between the individuals in the control group and those in the treatment group. Randomization eliminates just the systematic differences between the groups.

### Statistical significance

An observed effect of a size that would rarely occur by chance is called **statistically significant**.

The difference between the average number of pain episodes for subjects in the hydroxyurea group and the average for the control group was "highly statistically significant." That means that a difference of this size would almost never happen just by chance. We do indeed have strong evidence that hydroxyurea beats a placebo in helping sickle-cell disease sufferers. You will often see the phrase "statistically significant" in reports of investigations in many fields of study. It tells you that the investigators found good "statistical" evidence for the effect they were seeking.

Of course, the actual results of an experiment are more important than the seal of approval given by statistical significance. The treatment group in the sickle-cell experiment had an average of 2.5 pain episodes per year as opposed to 4.5 per year in the control group. That's a big enough difference to be important to people with the disease. A difference of 2.5 versus 2.8 would be much less interesting even if it were statistically significant.

How large an observed effect must be in order to be regarded as statistically significant depends on the number of subjects involved. A relatively small effect—one that might not be regarded as practically important—can be statistically significant if the size of the study is large. Thus, in the sickle-cell experiment, an average of 2.50 pain episodes per year versus

2.51 per year in the control group could be statistically significant if the number of subjects involved is sufficiently large. For a very large number of subjects, the average number of pain episodes per year should be almost the same if differences are due only to chance. It is also true that a very large effect may not be statistically significant. If the number of subjects in an experiment is small, it may be possible to observe large effects simply by chance. We will discuss these issues more fully in Parts III and IV.

Thus, in assessing statistical significance, it is helpful to know the magnitude of the observed effect and the number of subjects. Perhaps a better term than "statistically significant" might be "statistically dissimilar."

## How to live with observational studies

Do children who are bullied suffer depression as adults? Do doctors discriminate against women in treating heart disease? Does texting while driving increase the risk of having an accident? These are cause-and-effect questions, so we reach for our favorite tool, the randomized comparative experiment. Sorry. We refuse to require children to be bullied. We can't use random digits to assign heart disease patients to be men or women. We are reluctant to require drivers to use cell phones in traffic because talking while driving may be risky.

The best data we have about these and many other cause-and-effect questions come from observational studies. We know that observation is a weak second-best to experiment, but good observational studies are far from worthless, and we will discuss this further in Chapter 15. What makes a good observational study?

First, good studies are **comparative** even when they are not experiments. We compare random samples of people who were bullied as children with those who were not bullied. We compare how doctors treat men and women patients. We might compare drivers talking on cell phones with the *same* drivers when they are not on the phone. We can often combine comparison with **matching** in creating a control group. To see the effects of taking a painkiller during pregnancy, we compare women who did so with women who did not. From a large pool of women who did not take the drug, we select individuals who match the drug group in age, education, number of children, and other lurking variables. We now have two groups that are similar in all these ways so that these lurking variables should not affect our comparison of the groups. However, if other important lurking variables, not measurable or not thought of, are present, they will affect the comparison, and confounding will still be present.

Matching does not entirely eliminate confounding. People who were bullied as children may have characteristics that increase susceptibility to victimization as well as independently increasing the risk of depression.

They are more likely to be female, have had concurrent emotional or mental health problems as a child, have parents who suffer from depression, or have experienced maltreatment at home as a child. Although matching can reduce some of these differences, direct comparison of rates of depression in young adults who were bullied as children and in young adults who were not bullied as children would still confound any effect of bullying with the effects of mental health issues in childhood, mental health issues of the parents, and maltreatment as a child. A good comparative study **measures and adjusts for confounding variables**. If we measure sex, the presence of mental health issues as a child, the presence of mental health issues in the parents, and aspects of the home environment, there are statistical techniques that reduce the effects of these variables on rates of depression so that (we hope) only the effect of bullying itself remains.

### EXAMPLE 6 Bullying and depression

A recent study in the United Kingdom examined data on 3898 participants in a large observational study for which they had information on both victimization by peers at age 13 and the presence of depression at age 18. The researchers also had information on lots of variables, not just the explanatory variable (bullying at age 13) and the response variable (presence of depression at age 18). The research article said:

*Compared with children who were not victimized those who were frequently victimized by peers had over a twofold increase in the odds of depression.... This association was slightly reduced when adjusting for confounders...*

That "adjusting for confounders" means that the final results were adjusted for differences between the two groups. Adjustment reduced the association between bullying at age 13 and depression at age 18, but still left a nearly twofold increase in the odds of depression.

Interestingly, the researchers go on to mention that the use of observational data does not allow them to conclude the associations are causal.

### EXAMPLE 7 Sex bias in treating heart disease?

Doctors are less likely to give aggressive treatment to women with symptoms of heart disease than to men with similar symptoms. Is this because doctors are sexist? Not necessarily. Women tend to develop heart problems much later than men so that female heart patients are older and often have other health problems. That might explain why doctors proceed more cautiously in treating them.

This is a case for a comparative study with statistical adjustments for the effects of confounding variables. There have been several such studies, and they produce conflicting results. Some show, in the words of one doctor, "When men and women are otherwise the same and the only difference is gender, you find that treatments are very similar." Other studies find that women are undertreated even after adjusting for differences between the female and male subjects.

As Example 7 suggests, statistical adjustment is complicated. Randomization creates groups that are similar in *all* variables known and unknown. Matching and adjustment, on the other hand, can't work with variables the researchers didn't think to measure. Even if you believe that the researchers thought of everything, you should be a bit skeptical about statistical adjustment. There's lots of room for cheating in deciding which variables to adjust for. And the "adjusted" conclusion is really something like this:

*If female heart disease patients were younger and healthier than they really are, and if male patients were older and less healthy than they really are, then the two groups would get the same medical care.*

This may be the best we can get, and we should thank statistics for making such wisdom possible. But we end up longing for the clarity of a good experiment.

## STATISTICS IN SUMMARY

### Chapter Specifics

- Statistical studies often try to show that changing one variable (the **explanatory variable**) causes changes in another variable (the **response variable**).
- In an **experiment**, we actually set the explanatory variables ourselves rather than just observe them.
- Observational studies and one-track experiments that simply apply a single treatment often fail to produce useful data because **confounding** with **lurking variables** makes it impossible to say what the effect of the treatment was.
- In a **randomized comparative experiment** we compare two or more treatments, use chance to decide which subjects get each treatment, and use enough subjects so that the effects of chance are small.

- Comparing two or more treatments **controls** lurking variables affecting all subjects, such as the **placebo effect**, because they act on all the treatment groups.
- Differences among the effects of the treatments so large that they would rarely happen just by chance are called **statistically significant**.
- Observational studies of cause-and-effect questions are more impressive if they **compare matched groups** and measure as many lurking variables as possible to allow **statistical adjustment**.


In Chapter 1 we saw that experiments are best suited for drawing conclusions about whether a treatment causes a change in a response. In this chapter, we learned that only well-designed experiments, in particular randomized comparative experiments, provide a sound basis for such conclusions. Statistically significant differences among the effects of treatments are the best available evidence that changing the explanatory variable really *causes* changes in the response.

When it is not possible to do an experiment, observational studies that measure as many lurking variables as possible and make statistical adjustments for their effects are sometimes used to answer cause-and-effect questions. However, they remain a weak second-best to well-designed experiments.

**CASE STUDY EVALUATED** Use what you have learned in this chapter to evaluate the Case Study that opened the chapter. Start by reviewing the information on page 93. Then answer each of the following questions in complete sentences. Be sure to communicate clearly enough for any of your classmates to understand what you are saying.

First, here are the results of the study. After one season, the biomass of plants in the plot receiving additional spring rain was approximately twice that in plots receiving the other treatments. This difference was statistically significant.

1. Is this study an experiment or an observational study?
2. Explain what the phrase "statistically significant" means.
3. What advantage is gained by randomly assigning the plots to the treatments?

**LaunchPad Online Resources**
macmillan learning

- The Snapshots video *Types of Studies* and the StatClips video *Types of Studies* both review the differences between experiments and observational studies.
- The Snapshots video *Introduction to Statistics* describes real-world situations for which knowledge of statistical ideas is important.

- The StatBoards video *Factors and Treatments* identifies subjects, factors, treatments, and response variables in additional experiments.
- The StatBoards video *Outlining an Experiment* provides additional examples of outlining an experiment using figures similar to those given in this chapter.

## CHECK THE BASICS

*For Exercise 5.1, see page 101.*

**5.2 Explanatory and response variables.** Does regular church attendance lengthen people's lives? One study of the effect of regular attendance at religious services gathered data from a random sample of 3617 adults. The researchers measured whether a person attended religious services regularly and length of life. Which of the following is true?

(a) In this study, length of life is the explanatory variable and regular attendance of religious services is the response variable.

(b) In this study, regular attendance of religious services is the explanatory variable and length of life is the response variable.

(c) In this study, the 3617 adults are the explanatory variable and the information they provided is the response variable.

(d) In this study, there are no explanatory and response variables because these data come from a survey.

**5.3 Observational study or experiment?** The study described in Exercise 5.2 is

(a) a randomized comparative experiment.

(b) an experiment, but not a randomized experiment.

(c) an observational study.

(d) neither an experiment nor an observational study but, instead, a sample survey.

**5.4 Lurking variables.** People who attend church or synagogue are less likely to smoke or be overweight than nonattenders. In the study described in Exercise 5.2,

(a) smoking is a lurking variable, but weight is not.

(b) weight is a lurking variable, but smoking is not.

(c) smoking and weight are both lurking variables.

(d) neither smoking nor weight is a lurking variable.

**5.5 Statistical significance.** In the study described in Exercise 5.2, researchers found that, by the end of the study, there was a statistically significant difference in the likelihood of dying between those who regularly attend religious services and nonattenders. Nonattenders were 25% more likely to have died by the end of the study. Statistical significance here means

(a) the size of the observed difference in the likelihood of dying is not likely to be due to chance.

(b) the size of the observed difference in the likelihood of dying is likely to be due to chance.

(c) the size of the observed difference in the likelihood of dying has a 25% chance of occurring.

(d) the size of the observed difference in the likelihood of dying has a 75% chance of occurring.

**5.6 Randomized comparative experiment?** For which of the following studies would it be possible to conduct a randomized comparative experiment?

(a) A study to determine if the month you were born in affects how long you will live.

(b) A study to determine if taking Tylenol dulls your emotions.

(c) A study to determine if a person's sex affects his or her salary.

(d) A study to determine if the wealth of parents affects the wealth of their children.

## CHAPTER 5 EXERCISES


**5.7 Exhaust is bad for your heart.** A *CNET News* article reported that the artery walls of people living within 100 meters of a highway thicken more than twice as fast as the average person's. Researchers used ultrasound to measure the carotid artery wall thickness of 1483 people living near freeways in the Los Angeles area. The artery wall thickness among those living within 100 meters of a highway increased by 5.5 micrometers (roughly 1/20th the thickness of a human hair) each year during the three-year study, which is more than twice the progression observed in participants who did not live within this distance of a highway.

(a) What are the explanatory and response variables?

(b) Explain carefully why this study is not an experiment.

(c) Explain why confounding prevents us from concluding that living near a highway is bad for your heart because it causes increased thickness in the carotid artery wall.


**5.8 Birth month and health.** A *Columbus Dispatch* article reported that researchers at the Columbia University Department of Medicine examined records for an incredible 1.75 million patients born between 1900 and 2000 who had been treated at Columbia University Medical Center. Using statistical analysis, the researchers found that for cardiovascular disease, those born in the fall (September through December) were more protected, while those born in winter and spring (January to June) had higher risk. And because so many lives are cut short due to cardiovascular diseases, being born in the autumn was actually associated with living longer than being born in the spring. Is this conclusion the result of an experiment? Why or why not? What are the explanatory and response variables?


**5.9 Weight-loss surgery and longer life.** An article in the *Washington Post* reported that, according to two large studies, obese people are significantly less likely to die prematurely if they undergo stomach surgery to lose weight. But people choose whether to have stomach surgery. Explain why this fact makes any conclusion about cause

and effect untrustworthy. Use the language of lurking variables and confounding in your explanation, and draw a picture like Figure 5.1 to illustrate your explanation.

**5.10 Is obesity contagious?** A study closely followed a large social network of 12,067 people for 32 years, from 1971 until 2003. The researchers found that when a person gains weight, close friends tend to gain weight, too. The researchers reported that obesity can spread from person to person, much like a virus.

Explain why the fact that, when a person gains weight, close friends also tend to gain weight does not necessarily mean that weight gains in a person cause weight gains in close friends. In particular, identify some lurking variables whose effect on weight gain may be confounded with the effect of weight gains in close friends. Draw a picture like Figure 5.1 to illustrate your explanation.

**5.11 Aspirin and heart attacks.** Can aspirin help prevent heart attacks? The Physicians' Health Study, a large medical experiment involving 22,000 male physicians, attempted to answer this question. One group of about 11,000 physicians took an aspirin every second day, while the rest took a placebo. After several years, the study found that subjects in the aspirin group had significantly fewer heart attacks than subjects in the placebo group.

(a) Identify the experimental subjects, the explanatory variable and the values it can take, and the response variable.

(b) Use a diagram to outline the design of the Physicians' Health Study. (When you outline the design of an experiment, be sure to indicate the size of the treatment groups and the response variable. The diagrams in Figures 5.2 and 5.3 are models.)

(c) What do you think the term "significantly" means in "significantly fewer heart attacks"?

**5.12 The pen is mightier than the keyboard.** Is longhand note-taking more effective for learning than taking notes on a laptop? Researchers at two universities studied this issue. In one of the studies, 65 students listened to five talks. Students were randomly assigned either a laptop or a notebook for purposes of taking notes. Assume that 33 students were assigned to use laptops and 32 longhand. Whether taking notes on a laptop or by hand in a notebook, students were instructed to use their normal note-taking strategy. Thirty minutes after the lectures, participants were tested with conceptual application questions based on the lectures. Those taking notes by hand performed better than those taking notes on a laptop. Why is instructing students to use their normal note-taking strategy a problem if the goal is to determine the effect on learning of note-taking on a laptop as compared to note-taking by hand?

**5.13 Neighborhood's effect on grades.** To study the effect of neighborhood on academic performance, 1000 families were given federal housing vouchers to move out of their low-income neighborhoods. No improvement in the academic performance of the children in the families was found one year after the move.

Explain clearly why the lack of improvement in academic performance after one year does not necessarily

mean that neighborhood does not affect academic performance. In particular, identify some lurking variables whose effect on academic performance may be confounded with the effect of neighborhood. Use a picture like Figure 5.1 to illustrate your explanation.

**5.14 The pen is mightier than the keyboard, continued.**

(a) Outline the design of Exercise 5.12 for the experiment to compare the two treatments (laptop note-taking and longhand note-taking) that students received for taking notes. When you outline the design of an experiment, be sure to indicate the size of the treatment groups and the response variable. The diagrams in Figures 5.2 and 5.3 are models.

(b) If you have access to statistical software, use it to carry out the randomization required by your design. Otherwise, use Table A, beginning at line 119, to do the randomization your design requires.

**5.15 Learning on the Web.** The discussion following Example 1 notes that the Nova Southeastern University study does not tell us much about Web versus classroom learning because the students who chose the Web version were much better prepared. Describe the design of an experiment to get better information.


**5.16 Do antioxidants prevent cancer?** People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in "antioxidants" such as vitamins A, C, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied this question with 864 people who were at risk for colon cancer. The subjects were divided into four groups: daily beta-carotene, daily vitamins C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.

(a) What are the explanatory and response variables in this experiment?

(b) Outline the design of the experiment. (The diagrams in Figures 5.2 and 5.3 are models.)

(c) Assign labels to the 864 subjects. If you have access to statistical software, use it to choose the *first five* subjects for the beta-carotene group. Otherwise, use Table A, starting at line 118, to choose the *first five* subjects for the beta-carotene group.

(d) What does "no significant difference" mean in describing the outcome of the study?

(e) Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The results of the experiment suggest that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.

**5.17 Conserving energy.** Example 5 describes an experiment to learn whether providing households with electronic meters or with an app will reduce their electricity consumption. An executive of the electric company objects to including a control group. He says, "It would be cheaper to just compare electricity use last year [before the meter or app was provided] with consumption in the same period this year. If households use less electricity

this year, the meter or app must be working." Explain clearly why this design is inferior to that in Example 5.


**5.18 Improving Chicago's schools.** The National Science Foundation (NSF) paid for "systemic initiatives" to help cities reform their public education systems in ways that should help students learn better. Does this program work? The initiative in Chicago focused on improving the teaching of mathematics in high schools. The average scores of students on a standard test of math skills were higher after two years of the program in 51 out of 60 high schools in the city. Leaders of NSF said this was evidence that the Chicago program was succeeding. Critics said this doesn't say anything about the effect of the systemic initiative. Are these critics correct? Explain.


**5.19 Tylenol dulls emotions.** Will the same dose of Tylenol that stops the throbbing pain in your stubbed toe make you feel less joy at your sister's wedding? A *Columbus Dispatch* article reported on a study, conducted by researchers at The Ohio State University, of the effects of Tylenol on emotions. Half of the volunteers in the study were randomly assigned to take acetaminophen and the other half a placebo. After allowing time for the drug to take effect, the research team showed the college students 40 images that ranged from extremely unpleasant to extremely pleasant. On the one end were things such as close-up shots of malnourished children and city blocks destroyed in a war zone. On the other were images of children playing with kittens in a park, a big pile of money, and the faces of a couple in bed together. The intensity of the response to the images was measured for each subject by asking them to respond to the question, "To what extent is this picture positive or negative" using an 11-point scale from −5 (extremely negative) to 5 (extremely positive).

(a) What is the explanatory variable?

(b) What is the response variable, and what values does it take?

(c) Explain why the researchers gave half the volunteers a placebo rather than no treatment at all.


**5.20 Reducing health care spending.** Will people spend less on health care if their health insurance requires them to pay some part of the cost themselves? An experiment on this issue asked if the percentage of medical costs that is paid by health insurance has an effect both on the amount of medical care that people use and on their health. The treatments were four insurance plans. Each plan paid all medical costs above a ceiling. Below the ceiling, the plans paid 100%, 75%, 50%, or 0% of costs incurred.

(a) Outline the design of a randomized comparative experiment suitable for this study.

(b) Briefly describe the practical and ethical difficulties that might arise in such an experiment.

**5.21 Tylenol dulls emotions.** Consider again the Tylenol experiment of Exercise 5.19.

(a) Use a diagram to describe a randomized comparative experimental design for this experiment.

(b) Assume there were 20 subjects used in the experiment. Use software

or Table A, starting at line 120, to do the randomization required by your design.

**5.22 Treating drunk drivers.** Once a person has been convicted of drunk driving, one purpose of court-mandated treatment or punishment is to prevent future offenses of the same kind. Suggest three different treatments that a court might require. Then outline the design of an experiment to compare their effectiveness. Be sure to specify the response variables you will measure.

**5.23 Statistical significance.** A randomized comparative experiment examines whether the usual care of patients with chronic heart failure plus aerobic exercise training improves health status compared with the usual care alone. The researchers conclude that usual care plus exercise training confers modest but statistically significant improvements in self-reported health status compared with usual care without training. Explain what "statistically significant" means in the context of this experiment, as if you were speaking to a patient who knows no statistics.


**5.24 Statistical significance.** A study, mandated by Congress when it passed No Child Left Behind in 2002, evaluated 15 reading and math software products used by 9424 students in 132 schools across the country during the 2004–2005 school year. It is the largest study that has compared students who received the technology with those who did not, as measured by their scores on standardized tests. There were no statistically significant differences between students who used software and those who did not. Explain the meaning of "no statistically significant differences" in plain language.

**5.25 All the weight loss in half the time.** Some medical researchers suspect that 30 minutes of daily exercise will be just as effective as 60 minutes of daily exercise in reducing weight. You have available 50 heavy but healthy people who are willing to serve as subjects.

(a) Outline an appropriate design for the experiment.

(b) The names of the subjects appear below. If you have access to statistical software, use it to carry out the randomization required by your design. Otherwise, use Table A, beginning at line 131, to do the randomization required by your design. List the subjects you will assign to the group who will do 30 minutes of daily exercise.

| | | |
|---|---|---|
| Albright | Landgraf | Stagner |
| Ashmead | Lathrop | Stettler |
| Asihiro | Lefevre | Tan |
| Bai | Lewis | Tang |
| Bayer | Li | Thomas |
| Biller | Lim | Tirmenstein |
| Chen | Madaeni | Tompkins |
| Critchlow | Martin | Townsend |
| Davis | Patton | Turkmen |
| Dobmeier | Penzenik | Wang |
| Han | Powell | Westra |
| Hotait | Ren | Williams |
| Hu | Rodriguez | Winner |
| Josey | Samara | Yontz |
| Jung | Sanders | Yulovitch |
| Khalaf | Schneider | Zhang |
| Koster | Smith | |

**5.26 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 one or the other method. The study found that patients treated by the new method were significantly more likely to die within eight years.

(a) Further study of the data showed that this conclusion was wrong. The extra deaths among patients treated with the new method could be explained by lurking variables. What lurking variables might be confounded with a doctor's choice of surgical or nonsurgical treatment? For example, why might a doctor avoid assigning a patient to surgery?

(b) You have 300 prostate patients who are willing to serve as subjects in an experiment to compare the two methods. Use a diagram to outline the design of a randomized comparative experiment.

**5.27 Prayer and meditation.** You read in a magazine that "nonphysical treatments such as meditation and prayer have been shown to be effective in controlled scientific studies for such ailments as high blood pressure, insomnia, ulcers, and asthma." Explain in simple language what the article means by "controlled scientific studies" and why such studies might show that meditation and prayer are effective treatments for some medical problems.

**5.28 Exercise and bone loss.** Does regular exercise reduce bone loss in postmenopausal women? Here are two ways to study this question. Which design will produce more trustworthy data? Explain why.

1. A researcher finds 1000 postmenopausal women who exercise regularly. She matches each with a similar postmenopausal woman who does not exercise regularly, and she follows both groups for five years.

2. Another researcher finds 2000 postmenopausal women who are willing to participate in a study. She assigns 1000 of the women to a regular program of supervised exercise. The other 1000 continue their usual habits. The researcher follows both groups for five years.


**5.29 Safety of anesthetics.** The death rates of surgical patients differ for operations in which different anesthetics are used. An observational study found these death rates for four anesthetics:

| | | |
|---|---|---|
| Anesthetic: | Halothane | Pentothal |
| Death rate: | 1.7% | 1.7% |
| Anesthetic: | Cyclopropane | Ether |
| Death rate: | 3.4% | 1.9% |

This is *not* good evidence that cyclopropane is more dangerous than the other anesthetics. Suggest some lurking variables that may be confounded with the choice of anesthetic in surgery and that could explain the different death rates.

**5.30 Randomization at work.** To demonstrate how randomization reduces confounding, consider the following situation. A nutrition experimenter intends to compare the weight gain of prematurely born infants fed Diet A

with those fed Diet B. To do this, she will feed each diet to 10 prematurely born infants whose parents have enrolled them in the study. She has available 10 baby girls and 10 baby boys. The researcher is concerned that baby boys may respond more favorably to the diets, so if all the baby boys were fed Diet A, the experiment would be biased in favor of Diet A.

(a) Label the infants 00, 01, …, 19. Use Table A to assign 10 infants to Diet A. Or, if you have access to statistical software, use it to assign 10 infants to Diet A. Do this four times, using different parts of the table (or different runs of your software), and write down the four groups assigned to Diet A.

(b) The infants labeled 10, 11, 12, 13, 14, 15, 16, 17, 18, and 19 are the 10 baby boys. How many of these infants were in each of the four Diet A groups that you generated? What was the average number of baby boys assigned to Diet A? What does this suggest about the effect of randomization on reducing confounding?

**EXPLORING** THE WEB

*Follow the QR code to access exercises.*