CHAPTER 5

Experiments, Good and Bad

CASE STUDY When patients undergo surgery, the operating room is kept cool so that the physicians in heavy gowns will not be overheated. However, exposure to the cold can produce hypothermia in patients, increasing susceptibility to infection.

MedioImages/CORBIS

Medical researchers in Austria investigated whether maintaining a patient's body temperature close to normal by heating the patient during surgery decreases infection rates. Two hundred patients took part in the study and all were undergoing colon or rectal surgeries, which are associated with a high risk of infection. Patients were randomly assigned to either a normal or a hypothermic temperature group. Intravenous fluids were administered through a fluid warmer for both groups, but only the normal group had the warmers activated. A forced-air cover was used on the upper bodies of all patients, but it delivered air only to the normal group. In order to keep the surgeons and operating personnel from detecting the treatment a patient was receiving, shields and drapes were placed over all devices that would indicate the treatment. Patients did not know which treatment they received. The percentages of patients who developed infections were compared in order to determine differences in the treatments.

The EESEE story "Surgery in a Blanket" contains more information about this study. 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 are passive data collection. We observe, record, or measure, but we don't interfere. Experiments are active data production. Experimenters actively intervene by imposing some treatment in order to see what happens. All experiments and many observational studies are interested in the effect 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.

The vocabulary of experiments

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 Web sites 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.

EXAMPLE 2 The effects of day care

Should the government provide day care for low-income children? If day care helps these children stay in school and hold good jobs later in life, the government would save money by paying less welfare and collecting more taxes, so even those who are concerned only about the cost to the government might support day care programs. The Carolina Abecedarian Project (the name suggests learning the ABCs) has followed a group of children since 1972. The results show that good day care makes a big difference in later school and work.

The Abecedarian Project is an experiment in which the *subjects* are 111 people who in 1972 were healthy but low-income black infants in Chapel Hill, North Carolina. All the infants received nutritional supplements and help from social workers. Half, chosen at random, were also placed in an intensive preschool program. The experiment compares these two treatments. The *explanatory variable* is just "preschool, yes or no." There are many *response variables,* recorded over more than 30 years, including academic test scores, college attendance, and employment.

JUPITERIMAGES/BananaStock/Alamy

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 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 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 tests 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.

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.

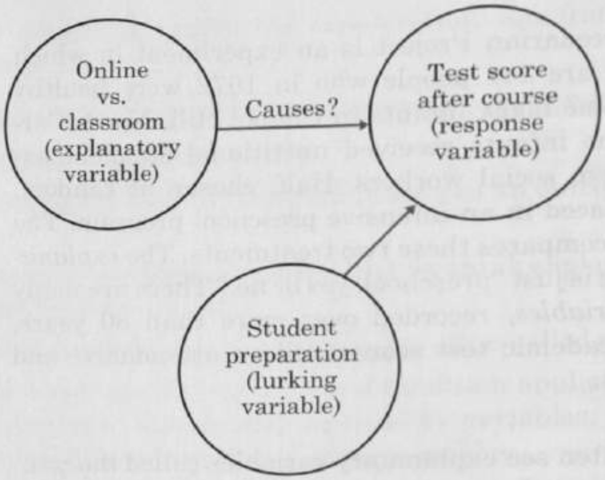
Online vs. classroom (explanatory variable)
Causes?
Test score after course (response variable)
Student preparation (lurking variable)

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).

In the Nova Southeastern 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, one in which only a single treatment was applied:

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.

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

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.

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. 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.

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

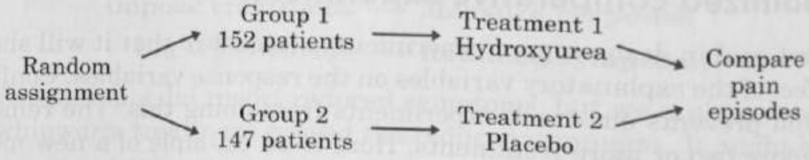

FIGURE 5.2 The design of a randomized comparative experiment to compare hydroxyurea with a placebo for treating sickle-cell anemia, for Example 4.

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 uses some of the techniques discussed in Chapter 2 for choosing a simple random sample. Label the 299 subjects 001 to 299, then 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 3 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. 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 a chart and information about monitoring their electricity use. The experiment compares these two approaches (meter, chart) 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 3 treatments. Figure 5.3 outlines the design.

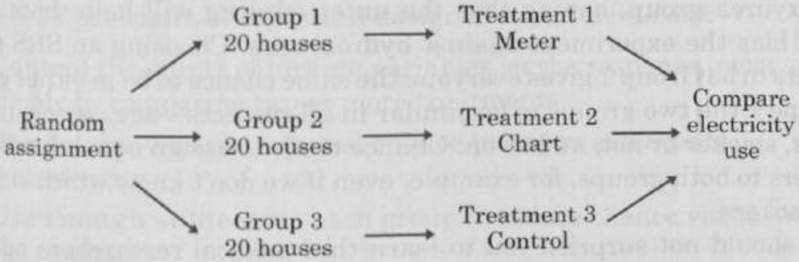

FIGURE 5.3 The design of a randomized comparative experiment to compare three programs to reduce electricity use by households, for Example 5.

To carry out the random assignment, 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 charts. The remaining 20 form the control group.

John Kelly/Getty Images

NOW IT'S YOUR TURN **5.1 Exercise and heart attacks.** Does regular exercise reduce the risk of a heart attack? To answer this question a researcher finds 4000 men over 40 who have not had heart attacks and are willing to participate in a study. She assigns 2000 of the men to a regular program of supervised exercise. The other 2000 continue their usual habits. The researcher follows both groups for 5 years. 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 in all respects before we apply the treatments.
- Comparative design ensures that influences other than the experimental treatments operate equally on all groups.
- 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.

It should 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 had 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, the 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 Google 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, most simply by comparing two or more treatments.
2. **Randomize**—use impersonal chance to assign subjects to treatments.
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 and control groups must be due to the effect of the drug. Even if both treatments are the same, there will always be some chance differences among the individuals who are assigned to the control or treatment. 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, against 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

Does regular church attendance lengthen people's lives? Do doctors discriminate against women in treating heart disease? Does talking on a cell phone 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 can't randomly assign people to attend church or not, because going to religious services is an expression of beliefs or their absence. 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. What makes a good observational study?

First, good studies are **comparative** even when they are not experiments. We compare random samples of people who do and who don't attend religious services regularly. 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 attend church or synagogue or mosque take better care of themselves than nonattenders. They are less likely to smoke, more likely to exercise, and less likely to be overweight. Although matching can reduce some of these differences, direct comparison of ages at death of attenders and nonattenders would still confound any effect of religion with the effects of healthy living. A good comparative study **measures and adjusts for confounding variables.** If we measure weight, smoking, and exercise, there are statistical techniques that reduce the effects of these variables on length of life so that (we hope) only the effect of religion itself remains.

EXAMPLE 6 Living longer through religion

One of the better studies of the effect of regular attendance at religious services gathered data from a random sample of 3617 adults. Random sampling is a good start. The researchers then measured lots of variables, not just the explanatory variable (religious activities) and the response variable (length of life). A news article said:

Churchgoers were more likely to be nonsmokers, physically active, and at their right weight. But even after health behaviors were taken into account, those not attending religious services regularly still were about 25% more likely to have died.

That "taken into account" means that the final results were adjusted for differences between the two groups. Adjustment reduced the advantage of religion but still left a large benefit.

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 tricky. 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

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. The remedy is to use a **randomized comparative experiment.** 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 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.** Statistically significant results from randomized comparative experiments are the best available evidence that changing the explanatory variable really *causes* changes in the response. 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.** Observational studies remain a weak second best to experiments for answering questions about causation.

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 81. You can also read the EESEE story "Surgery in a Blanket" for additional information. 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. The percentage of patients who developed postoperative infections was three times larger in the hypothermic group than in the normal group. 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 subjects to the treatments? ■

CHAPTER 5 EXERCISES

For Exercise 5.1, see page 88.

5.2 Treating breast cancer. What is the preferred treatment for breast cancer that is detected in its early stages? The most common treatment was once removal of the breast. It is now usual to remove only the tumor and nearby lymph nodes, followed by radiation. To study whether these treatments differ in their effectiveness, a medical team examines the records of 25 large hospitals and compares the survival times after surgery of all women who have had either treatment.

(a) What are the explanatory and response variables?

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

(c) Explain why confounding will prevent this study from discovering which treatment is more effective. (The current treatment was in fact recommended after a large randomized comparative experiment.)

FROM THE NEWS **5.3 Decline in math SAT scores.** A *New York Times* article reported that average math SAT scores for the high school class of 2007 dropped 3 points compared with scores in 2006. Officials of the College Board, the nonprofit organization that administers the SAT, suggested that increased numbers of students taking the SAT had contributed to the decline in scores. "The larger the population you get that takes the exam, it obviously knocks down the scores," said Gaston Caperton, the president of the College Board. Is this conclusion the result of an experiment? Why or why not? What are the explanatory and response variables?

FROM THE NEWS **5.4 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 it.

5.5 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.

5.6 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"?

5.7 Pleasant scents improve memory during sleep. A group of researchers at two universities plan to study whether delivering a pleasant scent during sleep affects memory. A group of 10 students at each university is selected to take part in the study. The students will play a version of the game "Concentration." This involves memorizing the location of card pairs on a computer screen. Upon learning the location of each pair, the students will receive a burst of rose scent in their noses, through a mask they will wear. The students will then sleep. During sleep the students will receive either bursts of rose scent or no scent. After waking, the students will be tested to see how many of the card locations they can recall. For simplicity, the researchers decide that all the students at one university will receive the burst of scent while sleeping and all the students at the other university will not receive the burst of scent while sleeping. Why is this a bad idea?

5.8 Neighborhood's effect on grades. To study the effect of neighborhood on academic performance, one thousand 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.

5.9 Pleasant scents improve memory during sleep, continued. **(a)** Outline a better design than that of Exercise 5.7 for an experiment to compare the two treatments (scent or no scent) that students received while sleeping. What do you suggest as a response variable? (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) Use Table A, starting at line 119, to do the randomization your design requires.

5.10 Learning on the Web. The discussion following Example 1 notes that the Nova Southeastern 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.

5.11 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 and use Table A, starting at line 118, to choose the *first 5* 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 experiment suggests that these variables, rather than the antioxidants, may be responsible for the observed benefits of fruits and vegetables.

5.12 Conserving energy. Example 5 describes an experiment to learn whether providing households with electronic meters or with charts 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 chart was provided] with consumption in the same period this year. If households use less electricity this year, the meter or chart must be working." Explain clearly why this design is inferior to that in Example 5.

5.13 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.

5.14 Tool abrasion and sharpening angle. A manufacturer of chisels is interested in determining how the angle at which the cutting edge is sharpened affects tool abrasion. To answer this question, engineers obtain 20 similar chisels. They sharpen five chisels at each of 22.5, 25, 27.5, and 30 degrees. Then they measure the amount of abrasion (rated on a scale of 1 to 10, with 10 being the worst) after cutting several mortises (square holes) in 3/4-inch hard maple boards.

(a) The individuals studied in this experiment are not people. What are they?

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

5.15 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.

5.16 Tool abrasion and sharpening angle. Use a diagram to describe a randomized comparative experimental design for the tool abrasion experiment of Exercise 5.14. Use Table A, starting at line 120, to do the randomization required by your design.

5.17 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.

5.18 Statistical significance. A randomized comparative experiment examines whether the drug memantine improves the cognition of patients with moderate to severe Alzheimer's disease. The subjects receive either memantine or a placebo for 24 weeks. The researchers conclude that on measures of cognition, the memantine group had significantly better outcomes than the placebo group. "Significant" in this conclusion means statistically significant. Explain what "statistically significant" means in the context of this experiment, as if you were speaking to a doctor who knows no statistics.

5.19 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.

5.20 Memantine and Alzheimer's disease. Some medical researchers suspect that the drug memantine improves the cognition of patients with moderate to severe Alzheimer's disease. You have available 50 people with moderate to severe Alzheimer's disease who are willing to serve as subjects.

(a) Outline an appropriate design for the experiment, taking the placebo effect into account.

(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 to whom you will give the drug.

Alomar
Andersen
Asihiro
Bennett
Bikalis
Chen
Clemente
Cranston
Critchlow
Curtis
Denman
Drake
Durr
Edwards
Farouk
Fratianna
George
Green
Guha
Guillen
Han
Hitchcock
Howard
Hruska
Imrani
James
Kaplan
Krushchev
Kumar
Lawless
Liang
Lin
Maldonado
Marsden
Mondesi
O'Brian
Ogle
Orosco
Powell
Rodriguez
Rosen
Solomon
Sroka
Tompkins
Townsend
Tullock
Underwood
Veras
Weimer
Zhang

5.21 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 8 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?

(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.

5.22 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.

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

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 5 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 5 years.

FROM THE NEWS **5.24 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	Cyclopropane	Ether
Death rate:	1.7%	1.7%	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.

5.25 Randomization at work. To demonstrate how randomization reduces confounding, consider the following situation. A nutrition experimenter intends to compare the weight gain of newly weaned male rats fed Diet A with that of rats fed Diet B. To do this, she will feed each diet to 10 rats. She has available 10 rats of genetic strain 1 and 10 of strain 2. Strain 1 is more vigorous, so if the 10 rats of strain 1 were fed Diet A, the effects of strain and diet would be confounded, and the experiment would be biased in favor of Diet A.

(a) Label the rats 00, 01, . . . , 19. Use Table A to assign 10 rats to Diet A. Do this four times, using different parts of the table, and write down the four groups assigned to Diet A.

(b) Unknown to the experimenter, the rats labeled 00, 02, 04, 06, 08, 10, 12, 14, 16, and 18 are the 10 strain 1 rats. How many of these rats were in each of the four Diet A groups that you generated? What was the average number of strain 1 rats assigned to Diet A?

EXPLORING THE WEB

5.26 Web-based exercise. Go to the *New England Journal of Medicine* Web site (http://content.nejm.org) and find the article "Initial Treatment of Aggressive Lymphoma with High-Dose Chemotherapy and Autologous Stem-Cell Support" by Milpied et al. in the March 25, 2004, issue. Was this a comparative study? Was randomization used? How many subjects took part? Were the results statistically significant? (If your institution does not have a subscription to the *New England Journal of Medicine,* you can find an abstract of the article at www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15044639&dopt=Abstract.)

You can find the latest medical research in the *Journal of the American Medical Association* (www.jama.ama-assn.org) and the *New England Journal of Medicine* (http://content.nejm.org). Many of the articles describe randomized comparative experiments, and even more use the language of statistical significance.

NOTES AND DATA SOURCES

Page 82 Example 1: Allan H. Schulman and Randi L. Sims, "Learning in an online format versus an in-class format: an experimental study," *T.H.E. Journal,* June 1999, pp. 54–56.

Page 82 Example 2: Details of the Carolina Abecedarian Project, including references to published work, can be found online at www.fpg.unc.edu/~abc.

Page 84 Example 3: R. W. Summers et al., "*Trichuris suis* therapy in Crohn's disease," *Gut,* 54 (2005), pp. 87–90.

Page 86 Example 4: Samuel Charache et al., "Effects of hydroxyurea on the frequency of painful crises in sickle cell anemia," *New England Journal of Medicine,* 332 (1995), pp. 1317–1322.

Page 89 H. Sacks, T. C. Chalmers, and H. Smith, Jr., "Randomized versus historical controls for clinical trials," *American Journal of Medicine,* 72 (1982), pp. 233–240.

Page 92 Example 6: Marilyn Ellis, "Attending church found factor in longer life," *USA Today,* August 9, 1999.

Page 92 Example 7: Dr. Daniel B. Mark, in Associated Press, "Age, not bias, may explain differences in treatment," *New York Times,* April 26, 1994. Dr. Mark was commenting on Daniel B. Mark et al., "Absence of sex bias in the referral of patients for cardiac catheterization," *New England Journal of Medicine,* 330 (1994), pp. 1101–1106. See the correspondence from D. Douglas Miller and Leslee Shaw, "Sex bias in the care of patients with cardiovascular disease," *New England Journal of Medicine,* 331 (1994), p. 883, for comments on a study with opposing results.

Page 94 Exercise 5.3: A. Finder, "Math and reading SAT scores drop," *New York Times,* August 28, 2007.

Page 94 Exercise 5.4: R. Stein, "Weight-loss surgery tied to a longer life," *Washington Post,* August 23, 2007.

Page 94 Exercise 5.6: Information about the Physicians' Health Study is available online at http://phs.bwh.harvard.edu/phs1.htm.

Page 96 Exercise 5.11: G. Kolata, "New study finds vitamins are not cancer preventers," *New York Times,* July 21, 1994. For the details, look in the *Journal of the American Medical Association* for the same date.

Page 96 Exercise 5.13: Letter to the editor by Stan Metzenberg, *Science,* 286 (1999), p. 2083.

Page 97 Exercise 5.15 is based on Christopher Anderson, "Measuring what works in health care," *Science,* 263 (1994), pp. 1080–1082.

Page 97 Exercise 5.19: A. R. Paley, "Software's benefits on tests in doubt," *Washington Post,* April 5, 2007.

Page 99 Exercise 5.24: L. E. Moses and F. Mosteller, "Safety of anesthetics," in J. M. Tanur et al. (eds.), *Statistics: A Guide to the Unknown,* 3rd edition, Wadsworth, 1989, pp. 15–24.