Experiments in the Real World

6

Photononstop/SuperStock

CASE STUDY Is caffeine dependence real? Researchers at the Johns Hopkins University School of Medicine wanted to determine if some individuals develop a serious addiction called caffeine dependence syndrome. Eleven volunteers were recruited who were diagnosed as caffeine dependent. For a two-day period, these volunteers were given a capsule that either contained their daily amount of caffeine or a fake (non-active) substance. Over another two-day period, at least one week after the first, the contents of the capsules received were switched. Whether the subjects first received the capsule containing caffeine or the capsule with the fake substance was determined by randomization. The subjects' diets were restricted during the study periods. All products with caffeine were prohibited, but to divert the subjects' attention from caffeine, products containing ingredients such as artificial sweeteners were also prohibited. Questionnaires assessing depression, mood, and the presence of certain physical symptoms were administered at the end of each two-day period. The subjects also completed a tapping task in which they were instructed to press a button 200 times as fast as they could. Finally, subjects were interviewed by a researcher, who did not know what was in the capsules the subjects had taken, to find other evidence of functional impairment. The Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) story "Is Caffeine Dependence Real?" contains more information about this study. EESEE stories are available in LaunchPad.

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.

Equal treatment for all

Probability samples are a big idea, but sampling in practice has difficulties that just using random samples doesn't solve. Randomized comparative experiments are also a big idea, but they don't solve all the difficulties

of experimenting. A sampler must know exactly what information she wants and must compose questions that extract that information from her sample. An experimenter must know exactly what treatments and responses he wants information about, and he must construct the apparatus needed to apply the treatments and measure the responses. This is what psychologists or medical researchers or engineers mean when they talk about "designing an experiment." We are concerned with the *statistical* side of designing experiments, ideas that apply to experiments in psychology, medicine, engineering, and other areas as well. Even at this general level, you should understand the practical problems that can prevent an experiment from producing useful data.

The logic of a randomized comparative experiment assumes that all the subjects are treated alike except for the treatments that the experiment is designed to compare. Any other unequal treatment can cause bias. Treating subjects exactly alike is hard to do.

EXAMPLE 1 Rats and rabbits

Rats and rabbits that are specially bred to be uniform in their inherited characteristics are the subjects in many experiments. However, animals, like people, can be quite sensitive to how they are treated. Here are two amusing examples of how unequal treatment can create bias.

Does a new breakfast cereal provide good nutrition? To find out, compare the weight gains of young rats fed the new product and rats fed a standard diet. The rats are randomly assigned to diets and are housed in large racks of cages. It turns out that rats in upper cages grow a bit faster than rats in bottom cages. If the experimenters put rats fed the new product at the top and those fed the standard diet below, the experiment is biased in favor of the new product. Solution: assign the rats to cages at random.

Another study looked at the effects of human affection on the cholesterol level of rabbits. All the rabbit subjects ate the same diet. Some (chosen at random) were regularly removed from their cages to have their furry heads scratched by friendly people. The rabbits who received affection had lower cholesterol. So affection for some but not other rabbits could bias an experiment in which the rabbits' cholesterol level is a response variable.

Double-blind experiments

Placebos "work." That bare fact means that medical studies must take special care to show that a new treatment is not just a placebo. Part of equal treatment for all is to be sure that the placebo effect operates on all subjects.

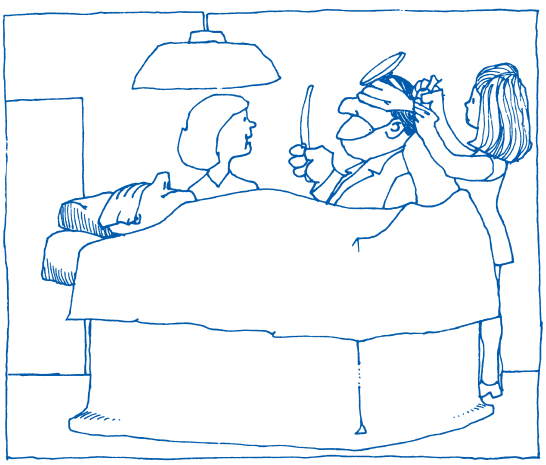
EXAMPLE 2 The powerful placebo

Want to help balding men keep their hair? Give them a placebo—one study found that 42% of balding men maintained or increased the amount of hair on their heads when they took a placebo. Another study told 13 people who were very sensitive to poison ivy that the stuff being rubbed on one arm was poison ivy. It was a placebo, but all 13 broke out in a rash. The stuff rubbed on the other arm really was poison ivy, but the subjects were told it was harmless—and only 2 of the 13 developed a rash.

When the ailment is vague and psychological, like depression, some experts think that about three-quarters of the effect of the most widely used drugs is just the placebo effect. Others disagree (see Web Exercise 6.31). The strength of the placebo effect in medical treatments is hard to pin down because it depends on the exact environment. How enthusiastic the doctor is seems to matter a lot. But "placebos work" is a good place to start when you think about planning medical experiments.

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

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



"Dr. Burns, are you sure this is what the statisticians call a double-blind experiment?"

Double-blind experiments

In a **double-blind experiment**, neither the subjects nor the people who work with them know which treatment each subject is receiving.

Until the study ends and the results are in, only the study's statistician knows for sure. Reports in medical journals regularly begin with words like these, from a study of a flu vaccine given as a nose spray: "This study was a randomized, double-blind, placebo-controlled trial. Participants were enrolled from 13 sites across the continental United States between mid-September and mid-November 1997." Doctors are expected to know what "randomized," "double-blind," and "placebo-controlled" mean. Now you also know.

Refusals, nonadherers, and dropouts

Sample surveys suffer from nonresponse due to failure to contact some people selected for the sample and the refusal of others to participate. Experiments with human subjects suffer from similar problems.

EXAMPLE 3 Minorities in clinical trials

Refusal to participate is a serious problem for medical experiments on treatments for major diseases such as cancer. As in the case of samples, bias can result if those who refuse are systematically different from those who cooperate.

Clinical trials are medical experiments involving human subjects. Minorities, women, the poor, and the elderly have long been underrepresented in clinical trials. In many cases, they weren't asked. The law now requires representation of women and minorities, and data show that most clinical trials now have fair representation. But refusals remain a problem. Minorities, especially blacks, are more likely to refuse to participate. The government's Office of Minority Health says, "Though recent studies have shown that African Americans have increasingly positive attitudes toward cancer medical research, several studies corroborate that they are still cynical about clinical trials. A major impediment for lack of participation is a lack of trust in the medical establishment." Some remedies for lack of trust are complete and clear information about the experiment, insurance coverage for experimental treatments, participation of black researchers, and cooperation with doctors and health organizations in black communities.

Subjects who participate but don't follow the experimental treatment, called **nonadherers**, can also cause bias. AIDS patients who participate in trials of a new drug sometimes take other treatments on their own, for example. In addition, some AIDS subjects have their medication tested and drop out or add other medications if they were not assigned to the new drug. This may bias the trial against the new drug.

Experiments that continue over an extended period of time also suffer **dropouts**, subjects who begin the experiment but do not complete it. If the reasons for dropping out are unrelated to the experimental treatments, no harm is done other than reducing the number of subjects. If subjects drop out because of their reaction to one of the treatments, bias can result.

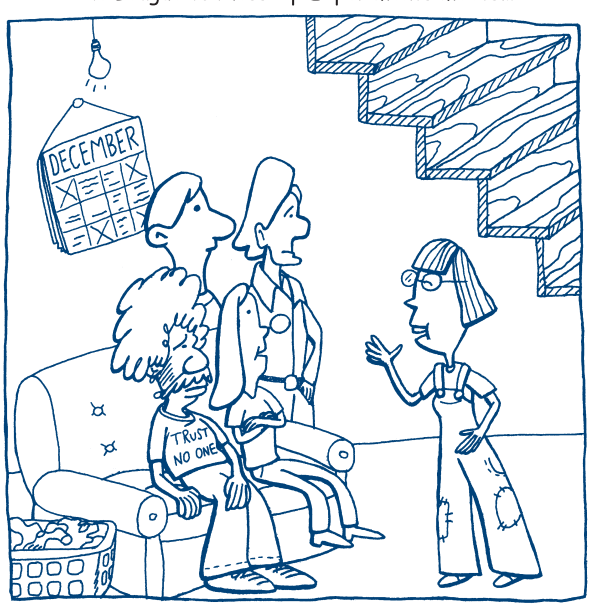
EXAMPLE 4 Dropouts in a medical study

Orlistat is a drug that may help reduce obesity by preventing absorption of fat from the foods we eat. As usual, the drug was compared with a placebo in a double-blind randomized trial. Here's what happened.

The subjects were 1187 obese subjects. They were given a placebo for four weeks, and the subjects who wouldn't take a pill regularly were dropped. This addressed the problem of nonadherers. There were 892 subjects left. These subjects were randomly assigned to Orlistat or a placebo, along with a weight-loss diet. After a year devoted to losing weight, 576 subjects were still participating. On the average, the Orlistat group lost 3.15 kilograms (about 7 pounds) more than the placebo group. The study continued for another year, now emphasizing maintaining the weight loss from the first year. At the end of the second year, 403 subjects were left. That's only 45% of the 892 who were randomized. Orlistat again beat the placebo, reducing the weight regained by an average of 2.25 kilograms (about 5 pounds).

Can we trust the results when so many subjects dropped out? The overall dropout rates were similar in the two groups: 54% of the subjects taking Orlistat and 57% of those in the placebo group dropped out. Were dropouts

The League to Mess Up Experiments meets...



"Agent B, you will scratch the heads of the lab rabbits. Agent Q, you will join a clinical trial and not take your pills. Agent K, you will sign up for an experiment, then drop out just before the end."

related to the treatments? Placebo subjects in weight-loss experiments often drop out because they aren't losing weight. This would bias the study against Orlistat because the subjects in the placebo group at the end may be those who could lose weight just by following a diet. The researchers looked carefully at the data available for subjects who dropped out. Dropouts from both groups had lost less weight than those who stayed, but careful statistical study suggested that there was little bias. Perhaps so, but the results aren't as clean as our first look at experiments promised.

Can we generalize?

A well-designed experiment tells us that changes in the explanatory variable cause changes in the response variable. More exactly, it tells us that this happened for specific subjects in the specific environment of this specific experiment. No doubt we had grander things in mind. We want to proclaim that our new method of teaching math does better for high school students in general or that our new drug beats a placebo for some broad class of patients. Can we generalize our conclusions from our little group of subjects to a wider population?

The first step is to be sure that our findings are *statistically significant*, that they are too strong to often occur just by chance. That's important, but it's a technical detail that the study's statistician can reassure us about. The serious threat is that the treatments, the subjects, or the environment of our experiment may not be realistic. Let's look at some examples.

EXAMPLE 5 Studying frustration

A psychologist wants to study the effects of failure and frustration on the relationships among members of a work team. She forms a team of students, brings them to the psychology laboratory, and has them play a game that requires teamwork. The game is rigged so that they lose regularly. The psychologist observes the students through a one-way window and notes the changes in their behavior during an evening of game playing.

Playing a game in a laboratory for small stakes, knowing that the session will soon be over, is a long way from working for months developing a new product that never works right and is finally abandoned by your company. Does the behavior of the students in the lab tell us much about the behavior of the team whose product failed?

In Example 5, the subjects (students who know they are subjects in an experiment), the treatment (a rigged game), and the environment (the psychology lab) are all unrealistic if the psychologist's goal is to reach conclusions about the effects of frustration on teamwork in the workplace.

Psychologists do their best to devise realistic experiments for studying human behavior, but lack of realism limits the ability to generalize beyond the environment and subjects in their study and, hence, the usefulness of some experiments in this area.

EXAMPLE 6 The effects of day care

Should the government provide day care for low-income preschool 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 Abecedarian Project is an experiment involving 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. Approximately half, chosen at random, were also placed in an intensive preschool program. The experiment compares these two treatments. Many response variables were recorded over more than 30 years, including academic test scores, college attendance, and employment.

This long and expensive experiment does show that intensive day care has substantial benefits in later life. The day care in the study was intensive indeed—lots of highly qualified staff, lots of parent participation, and detailed activities starting at a very young age, all costing about $11,000 per year for each child. It's unlikely that society will decide to offer such care to all low-income children, so the level of care in this experiment is somewhat unrealistic. The unanswered question is a big one: how good must day care be to really help children succeed in life?

EXAMPLE 7 Are subjects treated too well?

Surely medical experiments are realistic? After all, the subjects are real patients in real hospitals really being treated for real illnesses.

Even here, there are some questions. Patients participating in medical trials get better medical care than most other patients, even if they are in the placebo group. Their doctors are specialists doing research on their specific ailment. They are watched more carefully than other patients. They are more likely to take their pills regularly because they are constantly reminded to do so. Providing "equal treatment for all" except for the experimental and control therapies translates into "provide the best possible medical care for all." The result: ordinary patients may not do as well as the clinical trial subjects when the new

therapy comes into general use. It's likely that a therapy that beats a placebo in a clinical trial will beat it in ordinary medical care, but "cure rates" or other measures of success estimated from the trial may be optimistic.

Meta-analysis A single study of an important issue is rarely decisive. We often find several studies in different settings, with different designs, and of different quality. Can we combine their results to get an overall conclusion? That is the idea of "meta-analysis." Of course, differences among the studies prevent us from just lumping them together. Statisticians have more sophisticated ways of combining the results. Meta-analysis has been applied to issues ranging from the effect of secondhand smoke to whether coaching improves SAT scores.

When experiments are not fully realistic, statistical analysis of the experimental data cannot tell us how far the results will generalize. Experimenters generalizing from students in a lab to workers in the real world must argue based on their understanding of how people function, not based just on the data. It is even harder to generalize from rats in a lab to people in the real world. This is one reason a single experiment is rarely completely convincing, despite the compelling logic of experimental design. The true scope of a new finding must usually be explored by a number of experiments in various settings.

A convincing case that an experiment is sufficiently realistic to produce useful information is based not on statistics, but on the experimenter's knowledge of the subject matter of the experiment. The attention to detail required to avoid hidden bias also rests on subject-matter knowledge. Good experiments combine statistical principles with understanding of a specific field of study.

Experimental design in the real world

The experimental designs we have met all have the same pattern: divide the subjects at random into as many groups as there are treatments, then apply each treatment to one of the groups. These are *completely randomized* designs.

Completely randomized design

In a **completely randomized** experimental design, all the experimental subjects are allocated at random among all the treatments.

What is more, our examples to this point have had only a single explanatory variable (for example, drug versus placebo, classroom versus Web instruction). A completely randomized design can have any number of explanatory variables. Here is an example with two.

EXAMPLE 8 Can low-fat food labels lead to obesity?


Jamie Grill/Getty Images

What are the effects of low-fat food labels on food consumption? Do people eat more of a snack food when the food is labeled as low fat? The answer may depend both on whether the snack food is labeled low fat and whether the label includes serving-size information. An experiment investigated this question using university staff, graduate students, and undergraduate students at a large university as subjects. Over 10 late-afternoon sessions, all subjects viewed episodes of a 60-minute, made-for-television program in a theater on campus and were asked to rate the episodes. They were also told that because it was late in the afternoon, they would be given a cold 24-ounce bottle of water and a bag of granola from a respected campus restaurant called The Spice Box. They were told to enjoy as much or as little of it as they wanted. Each participant received 640 calories (160 grams) of granola in ziplock bags that were labeled with an attractive 3.25- ×- 4-inch color label. Depending on the condition randomly assigned to the subjects, the bags were labeled either "Regular Rocky Mountain Granola" or "Low-Fat Rocky Mountain Granola." Below this, the label indicated "Contains 1 Serving" or "Contains 2 Servings," or it provided no serving-size information. As participants left the theater, they were asked how many serving sizes they believed their package contained. Out of sight of the participants, the researchers also weighed each granola bag. Participants' statements about serving size and the actual weights of the granola bags are the response variables.

This experiment has two explanatory variables: fat content, with two levels, and serving size, with three levels. The six combinations of one level of each variable form six treatments. Figure 6.1 shows the layout of the treatments.

		Variable B Serving size		
		No information	1 serving	2 servings
Variable A Fat content	Regular	Treatment 1	Treatment 2	Treatment 3
	Low-fat	Treatment 4	Treatment 5	Treatment 6

Figure 6.1 The treatments in the experiment of Example 8. Combinations of two explanatory variables form six treatments.

NOW IT'S YOUR TURN

6.1 The perfect cup. One method for brewing coffee is the pour over. Ground coffee is placed in a filter-lined dripper (a metal or ceramic basket with holes in the bottom that fits over a carafe). Hot water is poured over the coffee and allowed to drip into the carafe. A coffee shop carries three brands of dripper, each brand differing in shape and the placement of holes through which the coffee drips. Some experts believe that moistening the filter with hot water before filling with grounds affects flavor, so the coffee shop plans the following experiment. For each brand of dripper, coffee is to be brewed both with the filter moistened and with the filter dry. A trained barista will brew two cups of coffee for each of the six combinations of dripper and wet/dry filter. A panel of trained tasters scores each cup for flavor on a scale of 1 to 10, with higher scores indicating better flavor.

What are the explanatory variables and the response variables for this experiment? Make a diagram like Figure 6.1 to describe the treatments. How many treatments are there? How many cups of coffee will be brewed?

Experimenters often want to study the combined effects of several variables simultaneously. The interaction of several factors can produce effects that could not be predicted from looking at the effect of each factor alone. Perhaps longer commercials increase interest in a product, and more commercials also increase interest, but if we both make a commercial longer and show it more often, viewers get annoyed and their interest in the product drops. An experiment similar to that in Example 8 will help us find out.

Matched pairs and block designs

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of control and randomization. However, completely randomized designs are often inferior to more elaborate statistical designs. In particular, matching the subjects in various ways can produce more precise results than simple randomization.

One common design that combines matching with randomization is the **matched pairs design**. A matched pairs design compares just two treatments. Choose pairs of subjects that are as closely matched as possible. Assign one of the treatments to each subject in a pair by tossing a coin or reading odd and even digits from Table A. Sometimes, each "pair" in a matched pairs design consists of just one subject, who gets both treatments together (for example, each on a different arm or leg) or one after

the other. Each subject serves as his or her own control. The *order* of the treatments can influence the subject's response, so we randomize the order for each subject, again by a coin toss.

EXAMPLE 9 Testing insect repellants

Consumers Reports describes a method for comparing the effectiveness of two insect repellants. The active ingredient in one is 15% Deet. The active ingredient in the other is oil of lemon eucalyptus. Repellants are tested on several volunteers. For each volunteer, the left arm is sprayed with one of the repellants and the right arm with the other. This is a matched pairs design in which each subject compares two insect repellants. To guard against the possibility that responses may depend on which arm is sprayed, which arm receives which repellant is determined randomly. Beginning 30 minutes after applying the repellants, once every hour, volunteers put each arm in separate 8-cubic-foot cages containing 200 disease-free female mosquitoes in need of a blood meal to lay their eggs. Volunteers leave their arms in the cages for five minutes. The repellant is considered to have failed if a volunteer is bitten two or more times in a five-minute session. The response is the number of one-hour sessions until a repellant fails.

Matched pairs designs use the principles of comparison of treatments and randomization. However, the randomization is not complete—we do not randomly assign all the subjects at once to the two treatments. Instead, we randomize only within each matched pair. This allows matching to reduce the effect of variation among the subjects. Matched pairs are an example of *block designs*.

Block design

A **block** is a group of experimental subjects that are known before the experiment to be similar in some way that is expected to affect the response to the treatments. In a **block design**, the random assignment of subjects to treatments is carried out separately within each block.

A block design combines the idea of creating equivalent treatment groups by matching with the principle of forming treatment groups at random. Blocks are another form of *control*. They control the effects of some outside variables by bringing those variables into the experiment to form the blocks. Here are some typical examples of block designs.

Hawthorne effect The *Hawthorne effect* is a term referring to the tendency of some people to work harder and perform better when they are participants in an experiment. Individuals may change their behavior due to the attention they are receiving from researchers rather than because of any manipulation of independent variables.

The effect was first described in the 1950s by researcher Henry A. Landsberger during his analysis of experiments conducted during the 1920s and 1930s at the Hawthorne works electric company.

The electric company had commissioned research to determine if there was a relationship between productivity and work environment.

The focus of the original studies was to determine if increasing or decreasing the amount of light workers received would have an effect on worker productivity. Employee productivity seemed to increase due to the changes but then decreased after the experiment was over. Researchers suggested that productivity increased due to attention from the research team and not because of changes to the experimental variables. Lansdberger defined the Hawthorne effect as a short-term improvement in performance caused by observing workers.

Later research into the Hawthorne effect has suggested that the original results may have been overstated. In 2009, researchers at the University of Chicago reanalyzed the original data and found that other factors also played a role in productivity and that the effect originally described was weak at best.

EXAMPLE 10 Men, women, and advertising

Women and men respond differently to advertising. An experiment to compare the effectiveness of three television commercials for the same product will want to look separately at the reactions of men and women, as well as assess the overall response to the ads.

A completely randomized design considers all subjects, both men and women, as a single pool. The randomization assigns subjects to three treatment groups without regard to their sex. This ignores the differences between men and women. A better design considers women and men separately. Randomly assign the women to three groups, one to view each commercial. Then separately assign the men at random to three groups. Figure 6.2 outlines this improved design.

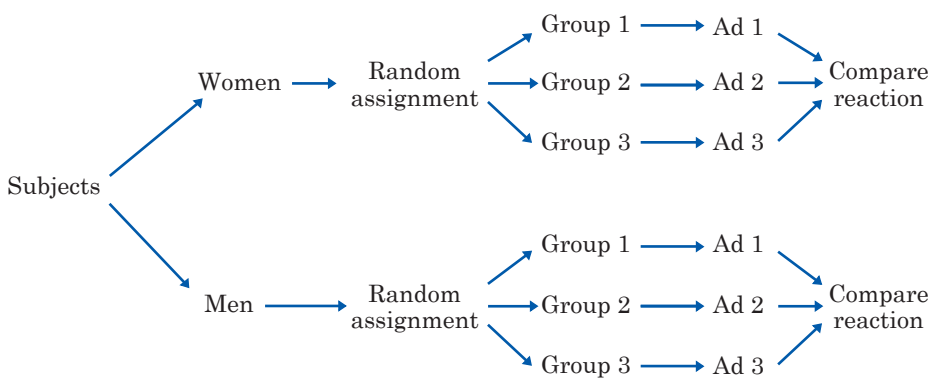


Figure 6.2 A block design to compare the effectiveness of three TV advertisements, Example 10. Female and male subjects form two blocks.

EXAMPLE 11 Comparing welfare systems

A social policy experiment will assess the effect on family income of several proposed new welfare systems and compare them with the present welfare system. Because the future income of a family is strongly related to its present income, the families who agree to participate are divided into blocks of similar income levels. The families in each block are then allocated at random among the welfare systems.

A block is a group of subjects formed before an experiment starts. We reserve the word "treatment" for a condition that we impose on the subjects. We don't speak of six treatments in Example 10 even though we can compare the responses of six groups of subjects formed by the two blocks (men, women) and the three commercials. Block designs are similar to stratified samples, which we discussed in Chapter 4. Blocks and strata both group similar individuals together. We use two different names only because the idea developed separately for sampling and experiments. The advantages of block designs are the same as the advantages of stratified samples. Blocks allow us to draw separate conclusions about each block—for example, about men and women in the advertising study in Example 10. Blocking also allows more precise overall conclusions because the systematic differences between men and women can be removed when we study the overall effects of the three commercials. The idea of blocking is an important additional principle of statistical design of experiments. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the experimental subjects. Randomization will then average out the effects of the remaining variation and allow an unbiased comparison of the treatments.

NOW IT'S YOUR TURN

6.2 Multiple-choice exams. A researcher was interested in whether the order of the answers to multiple-choice questions affects exam scores. He made three versions of an exam. Each version had the same questions and the same set of answers, but the order of the possible answers was different for each version. The three versions were given to students in two classes having different instructors. Each class had an enrollment of 75 students. The researcher was concerned that scores might also depend on instructor, so instructor was treated as a blocking variable. Use a diagram to outline a block design for this experiment. Use Figure 6.2 as a model.

Like the design of samples, the design of complex experiments is a job for experts. Now that we have seen a bit of what is involved, for the remainder of the text we will usually assume that most experiments were completely randomized.

STATISTICAL CONTROVERSIES

Is It or Isn't It a Placebo?

Natural supplements are big business: creatine and amino acid supplements to enhance athletic performance; green tea extract to boost the immune system; yohimbe bark to help your sex life; grapefruit extract and apple cider vinegar to support weight loss; white kidney bean extract to block carbs. Store shelves and websites are filled with exotic substances claiming to improve your health.

A therapy that has not been compared with a placebo in a randomized experiment may itself be just a placebo. In the United States, the law requires that new prescription drugs and new medical devices show their safety and effectiveness in randomized trials.

Cordelia Molloy/Science Source

What about those "natural remedies"? The law allows makers of herbs, vitamins, and dietary supplements to claim without any evidence that they are safe and will help "natural conditions." They can't claim to treat "diseases." Of course, the boundary between natural conditions and diseases is vague. Without any evidence whatsoever, we can claim that Dr. Moore's Old Indiana Extract promotes healthy hearts. But without clinical trials and an okay by the Food and Drug Administration (FDA), we can't claim that it reduces the risk of heart disease. No doubt lots of folks will think that "promotes healthy hearts" means the same thing as "reduces the risk of heart disease" when they see our advertisements. We also don't have to worry about what dose of Old Indiana Extract our pills contain or about what dose might actually be toxic.

Should the FDA require natural remedies to meet the same standards as prescription drugs? What does your statistical training tell you about claims not backed up by well-designed experiments? What about the fact that sometimes these natural remedies have real effects? Should that be sufficient for requiring FDA approval on natural remedies?

STATISTICS IN SUMMARY

Chapter Specifics

- Because the **placebo effect** is strong, **clinical trials** and other experiments with human subjects should be **double-blind** whenever this is possible.

- The double-blind method helps achieve a basic requirement of comparative experiments: **equal treatment for all subjects** except for the actual treatments the experiment is comparing.
- The most common weakness in experiments is that we can't **generalize** the conclusions widely. Some experiments apply unrealistic treatments, some use subjects from some special group such as college students, and all are performed at some specific place and time. We want to see similar experiments at other places and times confirm important findings.
- Many experiments use designs that are more complex than the basic **completely randomized design**, which divides all the subjects among all the treatments in one randomization. **Matched pairs designs** compare two treatments by giving one to each of a pair of similar subjects or by giving both to the same subject in random order. **Block designs** form blocks of similar subjects and assign treatments at random separately in each block.
- The big ideas of **randomization**, **control**, and **adequate numbers of subjects** remain the keys to convincing experiments.

In Chapter 5, we learned that well-designed randomized comparative experiments provide a sound basis for determining if a treatment causes changes in a response. In the real world, simple randomized comparative experiments don't solve all the difficulties of experimenting. The placebo effect and researchers' expectations can introduce biases that undermine our conclusions. Just as samples suffer from nonresponse, experiments suffer from uncooperative subjects. Some subjects refuse to participate; others drop out before the experiment is complete; others don't follow instructions, as when some subjects in a drug trial don't take their pills. More complex designs and techniques, some of which were discussed in this chapter, are used to overcome real-world difficulties. We must pay careful attention to every aspect of an experiment to ensure that the conclusions we make are valid. And when reading about the results of experiments, you should use the ideas provided in this chapter to assess the quality of the conclusions.

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 117. You can also read the EESEE story "Is Caffeine Dependence Real?" 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 number of subjects who showed withdrawal symptoms during the period in which they took capsules that did not contain caffeine and the magnitude of their symptoms were considered statistically significant.

1. Explain what the phrase "statistically significant" means.
2. Explain why the researchers gave subjects capsules with a fake substance rather than just having them take nothing during one of the periods.
3. What advantage is gained by having subjects take both a capsule with caffeine and a capsule with a fake substance rather than having some of the subjects just take a capsule with caffeine and the remaining subjects just take a capsule with a fake substance?

LaunchPad Online Resources
macmillan learning

- The Snapshot Video *Introduction to Statistics* describes real-world situations for which knowledge of statistical ideas is important.

CHECK THE BASICS

For Exercise 6.1, see page 126; for Exercise 6.2, see page 129.

6.3 Does meditation reduce anxiety? An experiment that claimed to show that meditation reduces anxiety proceeded as follows. The experimenter interviewed the subjects and rated their level of anxiety. The subjects were then randomly assigned to two groups. The experimenter taught one group how to meditate, and they meditated daily for a month. The other group was simply encouraged to relax more. At the end of the month, the experimenter interviewed all the subjects again and rated their anxiety level. The meditation group had a greater decrease in anxiety than the group told to relax more. These results might be biased because

(a) subjects should have been blinded to what treatment they received.

(b) the anxiety ratings at the end of the experiment should have been performed by someone blinded to which treatment a subject received.

(c) this is not a matched pairs design.

(d) the experimenter failed to use proper blocking.

6.4 Effects of TV advertising. What are the effects of repeated exposure to an advertising message? The answer may depend on both the length of the ad and how often it is repeated. An experiment investigated this question using undergraduate students. All students viewed a 40-minute television program that included ads for a new smartphone. Some subjects saw a 30-second commercial; others, a 90-second commercial. The same commercial was shown either one, three, or five times during the program. After viewing, all the students answered questions about their recall

of the ad, their attitude toward the smartphone, and their intention to purchase it. In this experiment, the length of the commercial and the number of times it was shown are

(a) the responses.

(b) the blocking variables.

(c) lurking variables.

(d) the explanatory variables.

6.5 Effects of TV advertising. Which of the following is an important weakness of the experiment described in Exercise 6.4?

(a) This was not a matched pairs design.

(b) Because undergraduate students were used as subjects and knew what was going on, the results may not generalize to everyday television viewers.

(c) This was not a double-blind experiment.

(d) This experiment did not use a placebo.

6.6 Reducing smoking. The Community Intervention Trial for Smoking Cessation asked whether a community-wide advertising campaign would reduce smoking. The researchers located 11 pairs of communities, with each pair similar in location, size, economic status, and so on. One community in each pair was chosen at random to participate in the advertising campaign and the other was not. This is

(a) an observational study.

(b) a matched pairs experiment.

(c) a completely randomized experiment.

(d) a randomized block design.

6.7 Wine and heart health. To explore the effects of red wine on heart health, you recruit 100 volunteers. Half are to drink one glass of red wine a day with dinner for a month. The other half are to abstain from any alcohol for a month. The diets of all volunteers are otherwise the same. Women and men may respond differently to wine. Forty of the volunteers are women and 60 are men, so the researchers separately randomly assign half the women to the wine group and half the men to the wine group. The remaining volunteers are assigned to the no alcohol group. This is an example of

(a) a completely randomized design.

(b) a matched pairs design.

(c) a block design.

(d) an observational study.

CHAPTER 6 EXERCISES

6.8 Magic mushrooms. A *Washington Post* article reported that psilocybin, the active ingredient of "magic mushrooms," promoted a mystical experience in two-thirds of people who took it for the first time, according to a study published in the online journal *Psychopharmacology*. The authors of the article stated that their "double-blind study evaluated the acute and longer-term psychological effects of a high dose of psilocybin relative to a comparison compound administered under comfortable, supportive conditions." Explain to someone who knows no statistics what the term "double-blind" means here.

6.9 Do antidepressants help? A researcher studied the effect of an antidepressant on depression. He randomly assigned subjects with moderate levels of depression to two groups. One group received the antidepressant and the other a placebo. Subjects were blinded with respect to the treatment they received. After four weeks, the researcher interviewed all subjects and rated the change in their symptoms based on the comments of subjects during the interview. Critics said that the results were suspect because the ratings were not blind. Explain what this means and how lack of blindness could bias the reported results.

6.10 Treating acne. An article in a medical journal reports an experiment to see if pulsed laser dye therapy is effective in treating acne. The article describes the experiment as a "randomized, controlled, single-blinded, split-face clinical trial of a volunteer sample of 40 patients aged 13 years or older with facial acne conducted at an academic referral center from August 2002 to September 2003." A split-face clinical trial is one in which one side of the face is treated and one side is not. What do you think "single-blinded" means here? Why isn't a double-blind experiment possible?

6.11 Bright bike lights. Will requiring bicyclists to use bright, high-intensity xenon lights mounted on the front and rear of the bike reduce accidents with cars by making bikes more visible?

(a) Briefly discuss the design of an experiment to help answer this question. In particular, what response variables will you examine?

(b) Suppose your experiment demonstrates that using high-intensity xenon lights reduces accidents. What concerns might you have about whether your experimental results will reduce accidents with cars if all bicyclists are required to use such lights? (*Hint*: To help you answer this question, consider the following example. A 1980 report by the Highway Traffic Safety Administration found that adding a center brake light to cars reduced rear-end collisions by as much as 50%. These findings were the result of a randomized comparative experiment. As a result, center brake lights have been required on all cars sold since 1986. Ten years later, the Insurance Institute found only a 5% reduction in rear-end collisions. Apparently, when the study was originally carried out, center brake lights were unusual and caught the eye of following drivers. By 1996, center brake lights were common and no longer captured attention.)

6.12 A high-fat diet prevents obesity? A *Science News* article reported that according to a study conducted by researchers at Hebrew University of Jerusalem, a high-fat diet could reset the metabolism and prevent obesity. In the study, for 18 weeks, researchers fed a group of mice a high-fat diet on a fixed schedule (eating at the same time and for the same length of time every day). They compared these mice to three control groups: one that ate a low-fat diet on a fixed schedule, one that ate an unscheduled low-fat diet (in the quantity and frequency of its choosing), and one that ate an unscheduled high-fat diet. All four groups of mice

gained weight throughout the experiment. However, the mice on the scheduled high-fat diet had a lower final body weight than the mice eating an unscheduled high-fat diet. Surprisingly, the mice on the scheduled high-fat diet also had a lower final body weight than the mice that ate an unscheduled low-fat diet, even though both groups consumed the same amount of calories. In addition, the mice on the scheduled high-fat diet exhibited a unique metabolic state in which the fats they ingested were not stored, but rather utilized for energy at times when no food was available, such as between meals. Briefly discuss the questions that arise in using this experiment to decide the benefits of a scheduled high-fat diet for humans.

6.13 Blood-chilling and strokes. A *Science News* article reported a study of the effect of cooling the blood of stroke patients on the extent of recovery 90 days after the stroke. Researchers randomly assigned 58 severe-stroke patients to receive either tPA (the standard treatment for stroke) or tPA plus blood-chilling. Regulators overseeing the study required a one-hour delay from the point at which tPA was given before cooling could be started. The researchers found no significant difference in the effects of the two treatments on recovery. Researchers also noted that the recovery rate for both groups was worse than the average seen in stroke patients nationwide but were not concerned. Why were they unconcerned?

6.14 Beating sunburn with broccoli. Some recent studies suggest that compounds in broccoli may be helpful in combating the effects of overexposure to ultraviolet radiation. Based on these studies, we hope to show that a cream consisting of a broccoli extract reduces sunburn pain. Sixty patients suffering from severe sunburn and needing pain relief are available. We will apply the cream to the sunburn of each patient and ask them an hour later, "About what percent of pain relief did you experience?"

(a) Why should we not simply apply the cream to all 60 patients and record the responses?

(b) Outline the design of an experiment to compare the cream's effectiveness with that of an over-the-counter product for sunburn relief and of a placebo.

(c) Should patients be told which remedy they are receiving? How might this knowledge affect their reactions?

(d) If patients are not told which treatment they are receiving, but the researchers assessing the effect of the treatment know, the experiment is single-blind. Should this experiment be double-blind? Explain.

6.15 Testing a natural remedy. The statistical controversy presented in this chapter discusses issues surrounding the efficacy of natural remedies. The National Institutes of Health at last began sponsoring proper clinical trials of some natural remedies. In one study at Duke University, 330 patients with mild depression were enrolled in a trial to compare Saint-John's-wort with a placebo and with Zoloft, a common prescription drug for depression. The Beck Depression Inventory is a

common instrument that rates the severity of depression on a 0 to 3 scale.

(a) What would you use as the response variable to measure change in depression after treatment?

(b) Outline the design of a completely randomized clinical trial for this study.

(c) What other precautions would you take in this trial?

6.16 The placebo effect. A survey of physicians found that some doctors give a placebo to a patient who complains of pain for which the physician can find no cause. If the patient's pain improves, these doctors conclude that it had no physical basis. The medical school researchers who conducted the survey claimed that these doctors do not understand the placebo effect. Why?

6.17 The best painkiller for children. A *Washington Post* article reported a study comparing the effectiveness of three common painkillers for children. Three hundred children, aged 6 to 17, were randomly assigned to three groups. Group A received a standard dose of ibuprofen. Group B received a standard dose of acetaminophen. Group C received a standard dose of codeine. The youngsters rated their pain on a 100-point scale before and after taking the medicine.

(a) Outline the design of this experiment. You do not need to do the randomization that your design requires.

(b) You read that "the children and physicians were blinded" during the study. What does this mean?

(c) You also read that there was a significantly greater decrease in pain ratings for Group A than for Groups B and C, but there was no significant difference in the decrease of pain ratings for Groups B and C. What does this mean? What does this finding lead you to conclude about the use of ibuprofen as a painkiller?

6.18 Flu shots. A *New York Times* article reported a study that investigated whether giving flu shots to schoolchildren protects a whole community from the disease. Researchers in Canada recruited 49 remote Hutterite farming colonies in western Canada for the study. In 25 of the colonies, all children aged 3 to 15 received flu shots in late 2008; in the 24 other colonies, they received a placebo. Which colonies received flu shots and which received the placebo was determined by randomization, and the colonies did not know whether they received the flu shots or the placebo. The researchers recorded the percentage of all children and adults in each colony who had laboratory-confirmed flu over the ensuing winter and spring.

(a) Outline the design of this experiment. You do not need to do the randomization that your design requires.

(b) The placebo was actually the hepatitis A vaccine, and "hepatitis was not studied, but to keep the investigators from knowing which colonies received flu vaccine, they had to offer placebo shots, and hepatitis shots do some good while sterile water injections do not." In addition, the article mentions that the colonies were studied "without the investigators being subconsciously biased by knowing which received the placebo." Why was it important that investigators not

be subconsciously biased by knowing which received the placebo?

(c) By June 2009, more than 10% of all the adults and children in colonies that received the placebo had had laboratory-confirmed seasonal flu. Less than 5% of those in the colonies that received flu shots had. This difference was statistically significant. Explain to someone who knows no statistics what "statistically significant" means in this context.

6.19 Ibuprofen and atherosclerosis. The theory of atherosclerosis (hardening and narrowing of the arteries) emphasizes the role of inflammation in the vascular walls. Because ibuprofen is known to possess a wide range of anti-inflammatory actions, it was hypothesized that it might help in the prevention of atherosclerotic lesion development. Both a low-cholesterol and a high-cholesterol diet were used, as the extent of atherosclerosis is also affected by diet. Thirty-two New Zealand rabbits served as subjects in the experiment and, after three months, the percentage of the surface covered by atherosclerotic plaques in a region of the aorta was evaluated. Although ibuprofen did suppress the expression of a gene thought to be related to atherosclerosis, it was not shown to have an effect on the extent of fat-induced atherosclerotic lesions.

(a) What are the individuals and the response variable in this experiment?

(b) How many explanatory variables are there? How many treatments? Use a diagram like Figure 6.1 to describe the treatments.

(c) Use a diagram to describe a completely randomized design for this experiment. (Don't actually do the randomization.)

6.20 Price change and fairness. A marketing researcher wishes to study what factors affect the perceived fairness of a change in the price of an item from its advertised price. In particular, does the type of change in price (an increase or decrease) and the source of the information about the change affect the perceived fairness? In an experiment, 20 subjects interested in purchasing a new rug are recruited. They are told that the price of a rug in a certain store was advertised at $500. Subjects are sent, one at a time, to the store, where they learn that the price has changed. Five subjects are told by a store clerk that the price has increased to $550. Five subjects learn that the price has increased to $550 from the price tag on the rug. Five subjects are told by a store clerk that the price has decreased to $450. Five subjects learn that the price has decreased to $450 from the price tag on the rug. After learning about the change in price, each subject is asked to rate the fairness of the change on a 10-point scale with 1 = "very unfair" to 10 = "very fair."

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

(b) Make a diagram like Figure 6.1 to describe the treatments. How many treatments are there?

(c) Explain why it is a bad idea to have the first five subjects learn from a store clerk that the price has increased to $550, the next five learn that the price has increased to $550 from the price tag on the rug, and so on.

Instead, the order in which subjects are sent to the store and which scenario they will encounter (type of change and source of information about the change) should be determined randomly. Why?

6.21 Liquid water enhancers. Bottled water, flavored and plain, is expected to become the largest segment of the liquid refreshment market by the end of the decade, surpassing traditional carbonated soft drinks. Kraft's MiO, a liquid water enhancer, comes in a variety of flavors, and a few drops added to water gives a zero-calorie, flavored water drink. You wonder if those who drink flavored water like the taste of MiO as well as they like the taste of a competing flavored water product that comes ready to drink. Describe a matched pairs design to answer this question. Be sure to include any blinding of your subjects. What is your response variable going to be?

6.22 Athletes take oxygen. We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this will speed their recovery. We might measure recovery from intense exertion as follows: Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Describe the design of two experiments to investigate the effect of inhaling oxygen during the rest period. One of the experiments is to be a completely randomized design and the other a matched pairs design in which each student serves as his or her own control. Twenty football players are available as subjects. In both experiments, carry out the randomization required by the design.

6.23 Font naturalness and perceived healthiness. Can the font used in a packaged product affect our perception of the product's healthiness? It was hypothesized that use of a natural font, which looks more handwritten and tends to be more slanted and curved, would lead to a higher perception of product healthiness than an unnatural font. Two fonts, Impact and Sketchflow Print, were used. These fonts were shown in a previous study to differ in their perceived naturalness, but otherwise were rated similarly on factors such as readability and likeability. Images of two identical packages, differing only in the font used, were available to be presented to the subjects. Participants read statements such as, "This product is healthy," "This product is wholesome," This product is natural," and "This product is organic." They then rated how much they agreed with the statements on a seven-point scale with "1" indicating strong agreement and "7" indicating strong disagreement. Each subject's responses were combined to create a perceived healthiness score. The researchers have 100 students available to serve as subjects.

(a) Outline a completely randomized design to learn the effect of font naturalness on perceived healthiness.

(b) Describe in detail the design of a matched pairs experiment, using the same 100 subjects, in which each subject serves as his or her own control.

6.24 Technology for teaching statistics. The Brigham Young University statistics department performed randomized comparative experiments to compare teaching methods. Response variables include students' final-exam scores and a measure of their attitude toward statistics. One study compared two levels of technology for large lectures: standard (overhead projectors and chalk) and multimedia. The individuals in the study were the eight lectures in a basic statistics course. There were four instructors, each of whom taught two lectures. Because lecturers differed, a block design was used with their lectures forming four blocks. Suppose the lectures and lecturers were as follows.

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

Use a diagram to outline a block design for this experiment. Figure 6.2 is a model.

6.25 Comparing weight-loss treatments. Twenty overweight females have agreed to participate in a study of the effectiveness of four weight-loss treatments: A, B, C, and D. The researcher first calculates how overweight each subject is by comparing the subject's actual weight with her "ideal" weight. The subjects and their excess weights in pounds are

Alexander	21	Murray	34
Barrasso	34	Nelson	28
Bayh	30	Pryor	30
Collins	25	Reed	30
Dodd	24	Sanders	27
Franken	25	Schumer	42
Hatch	33	Specter	33
Kerry	28	Tester	35
Leahy	32	Webb	29
McCain	39	Wyden	35

The response variable is the weight lost after eight weeks of treatment. Because a subject's excess weight will influence the response, a block design is appropriate.

(a) Arrange the subjects in order of increasing excess weight. Form five blocks of four subjects each by grouping the four least overweight, then the next four, and so on.

(b) Use Table A (or statistical software) to randomly assign the four subjects in each block to the four weight-loss treatments. Be sure to explain exactly how you used the table.

6.26 In the corn field. An agronomist (a specialist in crop production and soil chemistry) wants to compare the yield of four corn varieties. The field in which the experiment will be carried out increases in fertility from north to south. The agronomist therefore divides the field into 20 plots of equal size, arranged in five east-west rows of four plots each, and employs a block design with the rows of plots as the blocks.

(a) Draw a sketch of the field, divided into 20 plots. Label the rows Block 1 to Block 5.

(b) Do the randomization required by the block design. That is, randomly assign the four corn varieties A, B, C, and D to the four plots in each block. Mark on your sketch which variety is planted in each plot.

6.27 Better sleep? Is the time between when you go to bed and when you first wake up affected by the time you eat dinner and how much exercise you do during the day? Describe briefly the design of an experiment with two explanatory variables to investigate this question. Be sure to specify the treatments exactly and to tell how you will handle lurking variables such as amount of sleep the previous night.

6.28 Dunkin' Donuts versus Starbucks. Do consumers prefer the taste of a latte from Dunkin' Donuts or from Starbucks in a blind test in which neither latte is identified? Describe briefly the design of a matched pairs experiment to investigate this question.

6.29 What do you want to know? The previous two exercises illustrate the use of statistically designed experiments to answer questions that arise in everyday life. Select a question of interest to you that an experiment might answer and briefly discuss the design of an appropriate experiment.

6.30 Doctors and nurses. Nurse practitioners are nurses with advanced qualifications who often act much like primary-care physicians. An experiment assigned 1316 patients who had no regular source of medical care to either a doctor (510 patients) or a nurse practitioner (806 patients). All the patients had been diagnosed with asthma, diabetes, or high blood pressure before being assigned. The response variables included measures of the patients' health and of their satisfaction with their medical care after six months.

(a) Is the diagnosis (asthma, etc.) a treatment variable or a block? Why?

(b) Is the type of care (nurse or doctor) a treatment variable or a block? Why?

EXPLORING THE WEB

Follow the QR code to access exercises.