

Experiments and Observational Studies



Who gets good grades? And, more importantly, why? Is there something schools and parents could do to help weaker students improve their grades? Some people think they have an answer: music! No, not your iPod, but an instrument. In a study conducted at Mission Viejo High School, in California, researchers compared the scholastic performance of music students with that of non-music students. Guess what? The music students had a much higher overall grade point average than the non-music students, 3.59 to 2.91. Not only that: A whopping 16% of the music students had all A's compared with only 5% of the non-music students.

As a result of this study and others, many parent groups and educators pressed for expanded music programs in the nation's schools. They argued that the work ethic, discipline, and feeling of accomplishment fostered by learning to play an instrument also enhance a person's ability to succeed in school. They thought that involving more students in music would raise academic performance. What do you think? Does this study provide solid evidence? Or are there other possible explanations for the difference in grades? Is there any way to really prove such a conjecture?

Observational Studies

This research tried to show an association between music education and grades. But it wasn't a survey. Nor did it assign students to get music education. Instead, it simply observed students "in the wild," recording the choices they made and the outcome. Such studies are called **observational studies**. In observational studies, researchers don't *assign* choices; they simply observe them. In addition, this was a **retrospective study**, because researchers first identified subjects who studied music and then collected data on their past grades.

What's wrong with concluding that music education causes good grades? One high school during one academic year may not be representative of the

whole United States. That's true, but the real problem is that the claim that music study *caused* higher grades depends on there being *no other differences* between the groups that could account for the differences in grades, and studying music was not the *only* difference between the two groups of students.

We can think of lots of lurking variables that might cause the groups to perform differently. Students who study music may have better work habits to start with, and this makes them successful in both music and course work. Music students may have more parental support (someone had to pay for all those lessons), and that support may have enhanced their academic performance, too. Maybe they came from wealthier homes and had other advantages. Or it could be that smarter kids just like to play musical instruments.

For rare illnesses, it's not practical to draw a large enough sample to see many ill respondents, so the only option remaining is to develop retrospective data. For example, researchers can interview those who have become ill. The likely causes of both legionnaires' disease and HIV were initially identified from such retrospective studies of the small populations who were initially infected. But to confirm the causes, researchers needed laboratory-based experiments.

Observational studies are valuable for discovering trends and possible relationships. They are used widely in public health and marketing. Observational studies that try to discover variables related to rare outcomes, such as specific diseases, are often retrospective. They first identify people with the disease and then look into their history and heritage in search of things that may be related to their condition. But retrospective studies have a restricted view of the world because they are usually restricted to a small part of the entire population. **And because retrospective records are based on historical data, they can have errors.** (Do you recall *exactly* what you ate even yesterday? How about last Wednesday?)

A somewhat better approach is to observe individuals over time, recording the variables of interest and ultimately seeing how things turn out. For example, we might start by selecting young students who have not begun music lessons. We could then track their academic performance over several years, comparing those who later choose to study music with those who do not. **Identifying subjects in advance and collecting data as events unfold would make this a prospective study.**

Although an observational study may identify important variables related to the outcome we are interested in, there is no guarantee that we have found the right or the most important related variables. Students who choose to study an instrument might still differ from the others in some important way that we failed to observe. It may be this difference—whether we know what it is or not—rather than music itself that leads to better grades. It's just not possible for observational studies, whether prospective or retrospective, to demonstrate a causal relationship.

FOR EXAMPLE

Designing an observational study

In early 2007, a larger-than-usual number of cats and dogs developed kidney failure; many died. Initially, researchers didn't know why, so they used an observational study to investigate.

Question: Suppose you were called on to plan a study seeking the cause of this problem. Would your design be retrospective or prospective? Explain why.

I would use a retrospective observational study. Even though the incidence of disease was higher than usual, it was still rare. Surveying all pets would have been impractical. Instead, it makes sense to locate some who were sick and ask about their diets, exposure to toxins, and other possible causes.



Randomized, Comparative Experiments



Experimental design was advanced in the 19th century by work in psychophysics by Gustav Fechner (1801–1887), the founder of experimental psychology. Fechner designed ingenious experiments that exhibited many of the features of modern designed experiments. Fechner was careful to control for the effects of factors that might affect his results. For example, in his 1860 book *Elemente der Psychophysik* he cautioned readers to group experiment trials together to minimize the possible effects of time of day and fatigue.

An Experiment:

Manipulates the factor levels to create treatments.
Randomly assigns subjects to these treatment levels.
Compares the responses of the subject groups across treatment levels.

“He that leaves nothing to chance will do few things ill, but he will do very few things.”

—Lord Halifax
(1633–1695)

Is it *ever* possible to get convincing evidence of a cause-and-effect relationship? Well, yes it is, but we would have to take a different approach. We could take a group of third graders, randomly assign half to take music lessons, and forbid the other half to do so. Then we could compare their grades several years later. **This kind of study design is called an experiment.**

An experiment requires a **random assignment** of subjects to treatments. Only an experiment can justify a claim like “Music lessons cause higher grades.” Questions such as “Does taking vitamin C reduce the chance of getting a cold?” and “Does working with computers improve performance in Statistics class?” and “Is this drug a safe and effective treatment for that disease?” require a designed experiment to establish cause and effect.

Experiments study the relationship between two or more variables. An experimenter must identify at least one explanatory variable, called a **factor**, to manipulate and at least one **response variable** to measure. What distinguishes an experiment from other types of investigation is that the experimenter actively and deliberately manipulates the factors to control the details of the possible treatments, and assigns the subjects to those treatments *at random*. The experimenter then observes the response variable and *compares* responses for different groups of subjects who have been treated differently. For example, we might design an experiment to see whether the amount of sleep and exercise you get affects your performance.

The individuals on whom or which we experiment are known by a variety of terms. Humans who are experimented on are commonly called **subjects** or **participants**. Other individuals (rats, days, petri dishes of bacteria) are commonly referred to by the more generic term **experimental unit**. When we recruit subjects for our sleep deprivation experiment by advertising in Statistics class, we’ll probably have better luck if we invite them to be participants than if we advertise that we need experimental units.

The specific values that the experimenter chooses for a factor are called the **levels** of the factor. We might assign our participants to sleep for 4, 6, or 8 hours. Often there are several factors at a variety of levels. (Our subjects will also be assigned to a treadmill for 0 or 30 minutes.) The combination of specific levels from all the factors that an experimental unit receives is known as its **treatment**. (Our subjects could have any one of six different treatments—three sleep levels, each at two exercise levels.)

How should we assign our participants to these treatments? Some students prefer 4 hours of sleep, while others need 8. Some exercise regularly; others are couch potatoes. Should we let the students choose the treatments they’d prefer? No. That would not be a good idea. To have any hope of drawing a fair conclusion, we must assign our participants to their treatments *at random*.

It may be obvious to you that we shouldn’t let the students choose the treatment they’d prefer, but the need for random assignment is a lesson that was once hard for some to accept. For example, physicians might naturally prefer to assign patients to the therapy that they think best rather than have a random element such as a coin flip determine the treatment. But we’ve known for more than a century that for the results of an experiment to be valid, we must use deliberate randomization.

The Women’s Health Initiative is a major 15-year research program funded by the National Institutes of Health to address the most common causes of death, disability, and poor quality of life in older women. It consists of both an observational study with more than 93,000 participants and several randomized comparative experiments. The goals of this study include

- ▶ giving reliable estimates of the extent to which known risk factors predict heart disease, cancers, and fractures;

No drug can be sold in the United States without first showing, in a suitably designed experiment approved by the Food and Drug Administration (FDA), that it's safe and effective. The small print on the booklet that comes with many prescription drugs usually describes the outcomes of that experiment.

- ▶ identifying “new” risk factors for these and other diseases in women;
- ▶ comparing risk factors, presence of disease at the start of the study, and new occurrences of disease during the study across all study components; and
- ▶ creating a future resource to identify biological indicators of disease, especially substances and factors found in blood.

That is, the study seeks to identify possible risk factors and assess how serious they might be. It seeks to build up data that might be checked retrospectively as the women in the study continue to be followed. There would be no way to find out these things with an experiment because the task includes identifying new risk factors. If we don't know those risk factors, we could never control them as factors in an experiment.

By contrast, one of the clinical trials (randomized experiments) that received much press attention randomly assigned postmenopausal women to take either hormone replacement therapy or an inactive pill. The results published in 2002 and 2004 concluded that hormone replacement with estrogen carried increased risks of stroke.

FOR EXAMPLE

Determining the treatments and response variable

Recap: In 2007, deaths of a large number of pet dogs and cats were ultimately traced to contamination of some brands of pet food. The manufacturer now claims that the food is safe, but before it can be released, it must be tested.

Question: In an experiment to test whether the food is now safe for dogs to eat,¹ what would be the treatments and what would be the response variable?

The treatments would be ordinary-size portions of two dog foods: the new one from the company (the test food) and one that I was certain was safe (perhaps prepared in my kitchen or laboratory). The response would be a veterinarian's assessment of the health of the test animals.

The Four Principles of Experimental Design

AS

Video: An Industrial Experiment. Manufacturers often use designed experiments to help them perfect new products. Watch this video about one such experiment.

1. **Control.** We control sources of variation other than the factors we are testing by making conditions as similar as possible for all treatment groups. For human subjects, we try to treat them alike. However, there is always a question of degree and practicality. Controlling extraneous sources of variation reduces the variability of the responses, making it easier to detect differences among the treatment groups.

Making generalizations from the experiment to other levels of the controlled factor can be risky. For example, suppose we test two laundry detergents and carefully control the water temperature at 180°F. This would reduce the variation in our results due to water temperature, but what could we say about the detergents' performance in cold water? Not much. It would be hard to justify extrapolating the results to other temperatures.

Although we control both experimental factors and other sources of variation, we think of them very differently. We control a factor by assigning subjects to different factor levels because we want to see how the response will change at those different levels. We control other sources of variation to *prevent* them from changing and affecting the response variable.

¹ It may disturb you (as it does us) to think of deliberately putting dogs at risk in this experiment, but in fact that is what is done. The risk is borne by a small number of dogs so that the far larger population of dogs can be kept safe.

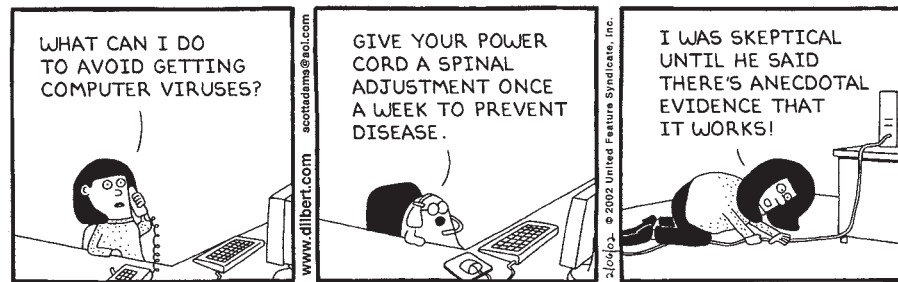


The deep insight that experiments should use random assignment is quite an old one. It can be attributed to the American philosopher and scientist C. S. Peirce in his experiments with J. Jastrow, published in 1885.

AS **Activity: The Three Rules of Experimental Design.** Watch an animated discussion of three rules of design.

AS **Activity: Perform an Experiment.** How well can you read pie charts and bar charts? Find out as you serve as the subject in your own experiment.

2. **Randomize.** As in sample surveys, **randomization** allows us to equalize the effects of unknown or uncontrollable sources of variation. It does not eliminate the effects of these sources, but it should spread them out across the treatment levels so that we can see past them. If experimental units were not assigned to treatments at random, we would not be able to use the powerful methods of Statistics to draw conclusions from an experiment. Assigning subjects to treatments at random reduces bias due to uncontrolled sources of variation. Randomization protects us even from effects we didn't know about. There's an adage that says "control what you can, and randomize the rest."
3. **Replicate.** Two kinds of replication show up in comparative experiments. First, we should apply each treatment to a number of subjects. Only with such replication can we estimate the variability of responses. If we have not assessed the variation, the experiment is not complete. The outcome of an experiment on a single subject is an anecdote, not data.



A second kind of replication shows up when the experimental units are not a representative sample from the population of interest. We may believe that what is true of the students in Psych 101 who volunteered for the sleep experiment is true of all humans, but we'll feel more confident if our results for the experiment are *replicated* in another part of the country, with people of different ages, and at different times of the year. **Replication of an entire experiment with the controlled sources of variation at different levels is an essential step in science.**

4. **Block.** The ability of randomizing to equalize variation across treatment groups works best in the long run. For example, if we're allocating players to two 6-player soccer teams from a pool of 12 children, we might do so at random to equalize the talent. But what if there were two 12-year-olds and ten 6-year-olds in the group? Randomizing may place both 12-year-olds on the same team. In the long run, if we did this over and over, it would all equalize. But wouldn't it be better to assign one 12-year-old to each group (at random) and five 6-year-olds to each team (at random)? By doing this, we would improve fairness in the short run. This approach makes the division more fair by recognizing the variation in *age* and allocating the players at random *within* each age level. When we do this, we call the variable *age* a **blocking variable**. The levels of *age* are called blocks.

Sometimes, attributes of the experimental units that we are not studying and that we can't control may nevertheless affect the outcomes of an experiment. If we group similar individuals together and then randomize within each of these **blocks**, we can remove much of the variability due to the difference among the blocks. Blocking is an important compromise between randomization and control. However, unlike the first three principles, blocking is not *required* in an experimental design.

FOR EXAMPLE

Control, randomize, and replicate

Recap: We're planning an experiment to see whether the new pet food is safe for dogs to eat. We'll feed some animals the new food and others a food known to be safe, comparing their health after a period of time.

Questions: In this experiment, how will you implement the principles of control, randomization, and replication?

I'd control the portion sizes eaten by the dogs. To reduce possible variability from factors other than the food, I'd standardize other aspects of their environments—housing the dogs in similar pens and ensuring that each got the same amount of water, exercise, play, and sleep time, for example. I might restrict the experiment to a single breed of dog and to adult dogs to further minimize variation.

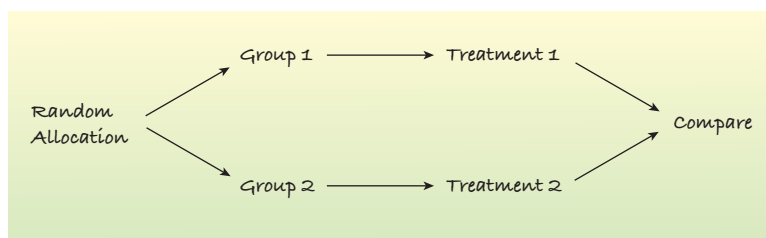
To equalize traits, pre-existing conditions, and other unknown influences, I would assign dogs to the two feed treatments randomly.

I would replicate by assigning more than one dog to each treatment to allow for variability among individual dogs. If I had the time and funding, I might replicate the entire experiment using, for example, a different breed of dog.



Diagrams

An experiment is carried out over time with specific actions occurring in a specified order. A diagram of the procedure can help in thinking about experiments.²



The diagram emphasizes the random allocation of subjects to treatment groups, the separate treatments applied to these groups, and the ultimate comparison of results. It's best to specify the responses that will be compared. A good way to start comparing results for the treatment groups is with boxplots.

STEP-BY-STEP EXAMPLE

Designing an Experiment



An ad for OptiGro plant fertilizer claims that with this product you will grow “juicier, tastier” tomatoes. You'd like to test this claim, and wonder whether you might be able to get by with half the specified dose. How can you set up an experiment to check out the claim?

Of course, you'll have to get some tomatoes, try growing some plants with the product and some without, and see what happens. But you'll need a clearer plan than that. How should you design your experiment?

² Diagrams of this sort were introduced by David Moore in his textbooks and are widely used.

A completely randomized experiment is the ideal simple design, just as a *simple random sample* is the ideal simple sample—and for many of the same reasons.

Let's work through the design, step by step. We'll design the simplest kind of experiment, a **completely randomized experiment in one factor**. Since this is a *design* for an experiment, most of the steps are part of the *Think* stage. The statements in the right column are the kinds of things you would need to say in *proposing* an experiment. You'd need to include them in the "methods" section of a report once the experiment is run.

Question: How would you design an experiment to test OptiGro fertilizer?



Plan State what you want to know.

I want to know whether tomato plants grown with OptiGro yield juicier, tastier tomatoes than plants raised in otherwise similar circumstances but without the fertilizer.

Response Specify the response variable.

I'll evaluate the juiciness and taste of the tomatoes by asking a panel of judges to rate them on a scale from 1 to 7 in juiciness and in taste.

Treatments Specify the factor levels and the treatments.

The factor is fertilizer, specifically OptiGro fertilizer. I'll grow tomatoes at three different factor levels: some with no fertilizer, some with half the specified amount of OptiGro, and some with the full dose of OptiGro. These are the three treatments.

Experimental Units Specify the experimental units.

I'll obtain 24 tomato plants of the same variety from a local garden store.

Experimental Design Observe the principles of design:

Control any sources of variability you know of and can control.

I'll locate the farm plots near each other so that the plants get similar amounts of sun and rain and experience similar temperatures. I will weed the plots equally and otherwise treat the plants alike.

Replicate results by placing more than one plant in each treatment group.

I'll use 8 plants in each treatment group.

Randomly assign experimental units to treatments, to equalize the effects of unknown or uncontrollable sources of variation.

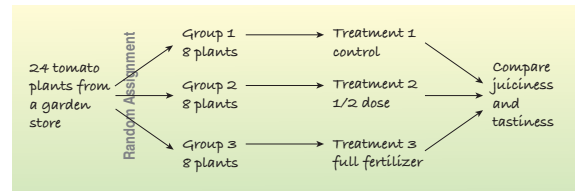
To randomly divide the plants into three groups, first I'll label the plants with numbers 00–23. I'll look at pairs of digits across a random number table. The first 8 plants identified (ignoring numbers 24–99 and any repeats) will go in Group 1, the next 8 in Group 2, and the remaining plants in Group 3.

Describe how the randomization will be accomplished.

Make a Picture A diagram of your design can help you think about it clearly.

Specify any other experiment details. You must give enough details so that another experimenter could exactly replicate your experiment. It's generally better to include details that might seem irrelevant than to leave out matters that could turn out to make a difference.

Specify how to measure the response.



I will grow the plants until the tomatoes are mature, as judged by reaching a standard color.

I'll harvest the tomatoes when ripe and store them for evaluation.

I'll set up a numerical scale of juiciness and one of tastiness for the taste testers. Several people will taste slices of tomato and rate them.

SHOW

Once you collect the data, you'll need to display them and compare the results for the three treatment groups.

I will display the results with side-by-side boxplots to compare the three treatment groups.

I will compare the means of the groups.

TELL

To answer the initial question, we ask whether the differences we observe in the means of the three groups are meaningful.

Because this is a randomized experiment, we can attribute significant differences to the treatments. To do this properly, we'll need methods from what is called "statistical inference," the subject of the rest of this book.

If the differences in taste and juiciness among the groups are greater than I would expect by knowing the usual variation among tomatoes, I may be able to conclude that these differences can be attributed to treatment with the fertilizer.

Does the Difference Make a Difference?

If the differences among the treatment groups are big enough, we'll attribute the differences to the treatments, but how can we decide whether the differences are big enough?

Would we expect the group means to be identical? Not really. Even if the treatment made no difference whatever, there would still be some variation. We assigned the tomato plants to treatments at random. But a different random assignment would have led to different results. Even a repeat of the *same* treatment on a different randomly assigned set of plants would lead to a different mean. The real question is whether the differences we observed are about as big as we might get just from the randomization alone, or whether they're bigger than that. If we decide that they're bigger, we'll attribute the differences to the treatments. In that case we say the differences are **statistically significant**.

A S

Activity: Graph the Data.

Do you think there's a significant difference in your perception of pie charts and bar charts? Explore the data from your plot perception experiment.

How will we decide if something is different enough to be considered statistically significant? Later chapters will offer methods to help answer that question, but to get some intuition, think about deciding whether a coin is fair. If we flip a fair coin 100 times, we expect, *on average*, to get 50 heads. Suppose we get 54 heads out of 100. That doesn't seem very surprising. It's well within the bounds of ordinary random fluctuations. What if we'd seen 94 heads? That's clearly outside the bounds. We'd be pretty sure that the coin flips were not random. But what about 74 heads? Is that far enough from 50% to arouse our suspicions? That's the sort of question we need to ask of our experiment results.

In Statistics terminology, 94 heads would be a statistically significant difference from 50, and 54 heads would not. Whether 74 is *statistically significant* or not would depend on the chance of getting 74 heads in 100 flips of a fair coin and on our tolerance for believing that rare events can happen to us.

Back at the tomato stand, we ask whether the differences we see among the treatment groups are the kind of differences we'd expect from randomization. A good way to get a feeling for that is to look at how much our results vary among plants that get the *same* treatment. Boxplots of our results by treatment group can give us a general idea.

For example, Figure 13.1 shows two pairs of boxplots whose centers differ by exactly the same amount. In the upper set, that difference appears to be larger than we'd expect just by chance. Why? Because the variation is quite small *within* treatment groups, so the larger difference *between* the groups is unlikely to be just from the randomization. In the bottom pair, that same difference between the centers looks less impressive. There the variation *within* each group swamps the difference *between* the two medians. We'd say the difference is statistically significant in the upper pair and not statistically significant in the lower pair.

In later chapters we'll see statistical tests that quantify this intuition. For now, the important point is that a difference is statistically significant if we don't believe that it's likely to have occurred only by chance.

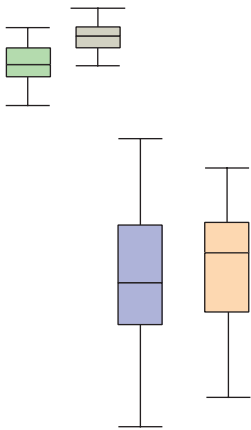


FIGURE 13.1

The boxplots in both pairs have centers the same distance apart, but when the spreads are large, the observed difference may be just from random fluctuation.



JUST CHECKING

- At one time, a method called “gastric freezing” was used to treat people with peptic ulcers. An inflatable bladder was inserted down the esophagus and into the stomach, and then a cold liquid was pumped into the bladder. Now you can find the following notice on the Internet site of a major insurance company:

[Our company] does not cover gastric freezing (intra-gastric hypothermia) for chronic peptic ulcer disease. . . .

Gastric freezing for chronic peptic ulcer disease is a non-surgical treatment which was popular about 20 years ago but now is seldom performed. It has been abandoned due to a high complication rate, only temporary improvement experienced by patients, and a lack of effectiveness when tested by double-blind, controlled clinical trials.

What did that “controlled clinical trial” (experiment) probably look like? (Don't worry about “double-blind”; we'll get to that soon.)

- | | |
|---|---|
| a) What was the factor in this experiment? | d) How did researchers decide which subjects received which treatment? |
| b) What was the response variable? | e) Were the results statistically significant? |
| c) What were the treatments? | |

Experiments and Samples

Both experiments and sample surveys use randomization to get unbiased data. But they do so in different ways and for different purposes. **Sample surveys try to estimate population parameters**, so the sample needs to be as representative of the population as possible. By contrast, **experiments try to assess the effects of treatments**. Experimental units are not always drawn randomly from the population. For example, a medical experiment may deal only with local patients who

have the disease under study. The randomization is in the assignment of their therapy. We want a sample to exhibit the diversity and variability of the population, but for an experiment the more homogeneous the subjects the more easily we'll spot differences in the effects of the treatments.



Experiments are rarely performed on random samples from a population. Don't describe the subjects in an experiment as a random sample unless they really are. More likely, the randomization was in assigning subjects to treatments.

Unless the experimental units are chosen from the population at random, you should be cautious about generalizing experiment results to larger populations until the experiment has been repeated under different circumstances. Results become more persuasive if they remain the same in completely different settings, such as in a different season, in a different country, or for a different species, to name a few.

Even without choosing experimental units from a population at random, experiments can draw stronger conclusions than surveys. By looking only at the differences across treatment groups, experiments cancel out many sources of bias. For example, the entire pool of subjects may be biased and not representative of the population. (College students may need more sleep, on average, than the general population.) When we assign subjects randomly to treatment groups, all the groups are still biased, but *in the same way*. When we consider the differences in their responses, these biases cancel out, allowing us to see the *differences* due to treatment effects more clearly.

Control Treatments

A S

Activity: Control Groups in Experiments. Is a control group really necessary?

Suppose you wanted to test a \$300 piece of software designed to shorten download times. You could just try it on several files and record the download times, but you probably want to *compare* the speed with what would happen *without* the software installed. Such a baseline measurement is called a **control treatment**, and the experimental units to whom it is applied are called a **control group**.

This is a use of the word “control” in an entirely different context. Previously, we controlled extraneous sources of variation by keeping them constant. Here, we use a control treatment as another *level* of the factor in order to compare the treatment results to a situation in which “nothing happens.” That’s what we did in the tomato experiment when we used no fertilizer on the 8 tomatoes in Group 1.

Blinding

Humans are notoriously susceptible to errors in judgment.³ All of us. When we know what treatment was assigned, it’s difficult not to let that knowledge influence our assessment of the response, even when we try to be careful.

Suppose you were trying to advise your school on which brand of cola to stock in the school’s vending machines. You set up an experiment to see which of the three competing brands students prefer (or whether they can tell the difference at all). But people have brand loyalties. You probably prefer one brand already. So if you knew which brand you were tasting, it might influence your rating. To avoid this problem, it would be better to disguise the brands as much as possible. This strategy is called **blinding** the participants to the treatment.⁴

But it isn’t just the subjects who should be blind. Experimenters themselves often subconsciously behave in ways that favor what they believe. Even technicians may treat plants or test animals differently if, for example, they expect them to die. An animal that starts doing a little better than others by showing an increased appetite may get fed a bit more than the experimental protocol specifies.

³ For example, here we are in Chapter 13 and you’re still reading the footnotes.

⁴ C. S. Peirce, in the same 1885 work in which he introduced randomization, also recommended blinding.

Blinding by Misleading

Social science experiments can sometimes blind subjects by misleading them about the purpose of a study. One of the authors participated as an undergraduate volunteer in a (now infamous) psychology experiment using such a blinding method. The subjects were told that the experiment was about three-dimensional spatial perception and were assigned to draw a model of a horse. While they were busy drawing, a loud noise and then groaning were heard coming from the room next door. The *real* purpose of the experiment was to see how people reacted to the apparent disaster. The experimenters wanted to see whether the social pressure of being in groups made people react to the disaster differently. Subjects had been randomly assigned to draw either in groups or alone; that was the treatment. The experimenter had no interest in how well the subjects could draw the horse, but the subjects were blinded to the treatment because they were misled.

People are so good at picking up subtle cues about treatments that the best (in fact, the *only*) defense against such biases in experiments on human subjects is to keep *anyone* who could affect the outcome or the measurement of the response from knowing which subjects have been assigned to which treatments. So, not only should your cola-tasting subjects be blinded, but also *you*, as the experimenter, shouldn't know which drink is which, either—at least until you're ready to analyze the results.

There are two main classes of individuals who can affect the outcome of the experiment:

- ▶ those who could influence the results (the subjects, treatment administrators, or technicians)
- ▶ those who evaluate the results (judges, treating physicians, etc.)

When all the individuals in either one of these classes are blinded, an experiment is said to be **single-blind**. When everyone in *both* classes is blinded, we call the experiment **double-blind**. Even if several individuals in one class are blinded—for example, both the patients and the technicians who administer the treatment—the study would still be just single-blind. If only some of the individuals in a class are blind—for example, if subjects are not told of their treatment, but the administering technician is not

blind—there is a substantial risk that subjects can discern their treatment from subtle cues in the technician's behavior or that the technician might inadvertently treat subjects differently. Such experiments cannot be considered truly blind.

In our tomato experiment, we certainly don't want the people judging the taste to know which tomatoes got the fertilizer. That makes the experiment single-blind. We might also not want the people caring for the tomatoes to know which ones were being fertilized, in case they might treat them differently in other ways, too. We can accomplish this double-blinding by having some fake fertilizer for them to put on the other plants. Read on.

FOR EXAMPLE**Blinding**

Recap: In our experiment to see if the new pet food is now safe, we're feeding one group of dogs the new food and another group a food we know to be safe. Our response variable is the health of the animals as assessed by a veterinarian.

Questions: Should the vet be blinded? Why or why not? How would you do this? (Extra credit: Can this experiment be double-blind? Would that mean that the test animals wouldn't know what they were eating?)

Whenever the response variable involves judgment, it is a good idea to blind the evaluator to the treatments. The veterinarian should not be told which dogs ate which foods.

Extra credit: There is a need for double-blinding. In this case, the workers who care for and feed the animals should not be aware of which dogs are receiving which food. We'll need to make the "safe" food look as much like the "test" food as possible.

Placebos

AS **Activity: Blinded Experiments.** This narrated account of blinding isn't a placebo!

Often, simply applying *any* treatment can induce an improvement. Every parent knows the medicinal value of a kiss to make a toddler's scrape or bump stop hurting. Some of the improvement seen with a treatment—even an effective treatment—can be due simply to the act of treating. To separate these two effects, we can use a control treatment that mimics the treatment itself.

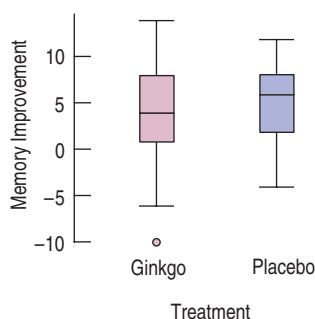
The placebo effect is stronger when placebo treatments are administered with authority or by a figure who appears to be an authority. “Doctors” in white coats generate a stronger effect than salespeople in polyester suits. But the placebo effect is not reduced much even when subjects know that the effect exists. People often suspect that they’ve gotten the placebo if nothing at all happens. So, recently, drug manufacturers have gone so far in making placebos realistic that they cause the same side effects as the drug being tested! Such “active placebos” usually induce a stronger placebo effect. When those side effects include loss of appetite or hair, the practice may raise ethical questions.

A “fake” treatment that looks just like the treatments being tested is called a **placebo**. Placebos are the best way to blind subjects from knowing whether they are receiving the treatment or not. One common version of a placebo in drug testing is a “sugar pill.” Especially when psychological attitude can affect the results, control group subjects treated with a placebo may show an improvement.

The fact is that subjects treated with a placebo sometimes improve. It’s not unusual for 20% or more of subjects given a placebo to report reduction in pain, improved movement, or greater alertness, or even to demonstrate improved health or performance. This **placebo effect** highlights both the importance of effective blinding and the importance of comparing treatments with a control. Placebo controls are so effective that you should use them as an essential tool for blinding whenever possible.

The best experiments are usually

- ▶ randomized.
- ▶ double-blind.
- ▶ comparative.
- ▶ placebo-controlled.



Does ginkgo biloba improve memory? Researchers investigated the purported memory-enhancing effect of ginkgo biloba tree extract (P. R. Solomon, F. Adams, A. Silver, J. Zimmer, R. De Veaux, “Ginkgo for Memory Enhancement. A Randomized Controlled Trial.” *JAMA* 288 [2002]: 835–840). In a randomized, comparative, double-blind, placebo-controlled study, they administered treatments to 230 elderly community members. One group received Ginkoba™ according to the manufacturer’s instructions. The other received a similar-looking placebo. Thirteen different tests of memory were administered before and after treatment. The placebo group showed greater improvement on 7 of the tests, the treatment group on the other 6. None showed any significant differences. Here are boxplots of one measure.



By permission of John L. Hart FLP and Creators Syndicate, Inc.

Blocking

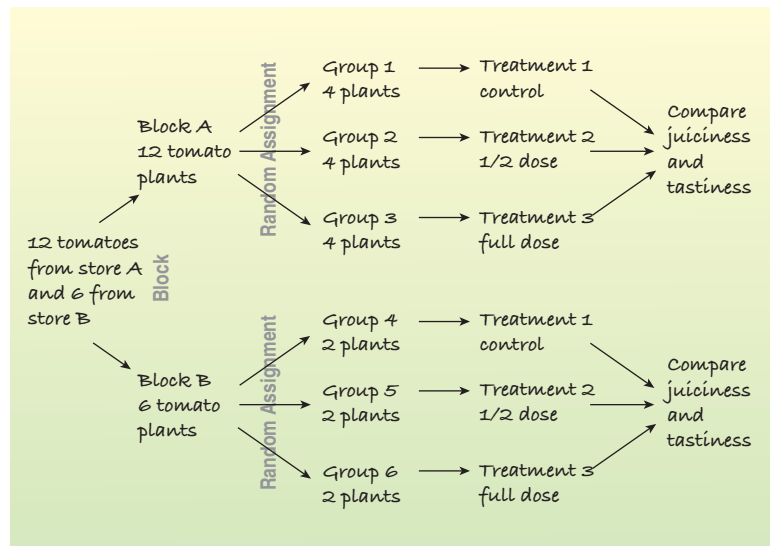
We wanted to use 18 tomato plants of the same variety for our experiment, but suppose the garden store had only 12 plants left. So we drove down to the nursery and bought 6 more plants of that variety. We worry that the tomato plants from the two stores are different somehow, and, in fact, they don’t really look the same.

How can we design the experiment so that the differences between the stores don’t mess up our attempts to see differences among fertilizer levels? We can’t measure the effect of a store the same way as we can the fertilizer because we can’t assign it as we would a factor in the experiment. You can’t tell a tomato what store to come from.

Because stores may vary in the care they give plants or in the sources of their seeds, the plants from either store are likely to be more like each other than they are like the plants from the other store. When groups of experimental units are similar, it's often a good idea to gather them together into **blocks**. By blocking, we isolate the variability attributable to the differences between the blocks, so that we can see the differences caused by the treatments more clearly. Here, we would define the plants from each store to be a block. The randomization is introduced when we randomly assign treatments within each block.

In a completely randomized design, each of the 18 plants would have an equal chance to land in each of the three treatment groups. But we realize that the store may have an effect. To isolate the store effect, we block on store by assigning the plants from each store to treatments at random. So we now have six treatment groups, three for each block. Within each block, we'll randomly assign the same number of plants to each of the three treatments. The experiment is still fair because each treatment is still applied (at random) to the same number of plants and to the same proportion from each store: 4 from store A and 2 from store B. Because the randomization occurs only within the blocks (plants from one store cannot be assigned to treatment groups for the other), we call this a **randomized block design**.

In effect, we conduct two parallel experiments, one for tomatoes from each store, and then combine the results. The picture tells the story:



In a retrospective or prospective study, subjects are sometimes paired because they are similar in ways *not* under study. **Matching** subjects in this way can reduce variation in much the same way as blocking. For example, a retrospective study of music education and grades might match each student who studies an instrument with someone of the same sex who is similar in family income but didn't study an instrument. When we compare grades of music students with those of non-music students, the matching would reduce the variation due to income and sex differences.

Blocking is the same idea for experiments as stratifying is for sampling. Both methods group together subjects that are similar and randomize within those groups as a way to remove unwanted variation. (But be careful to keep the terms straight. Don't say that we "stratify" an experiment or "block" a sample.) We use blocks to reduce variability so we can see the effects of the factors; we're not usually interested in studying the effects of the blocks themselves.

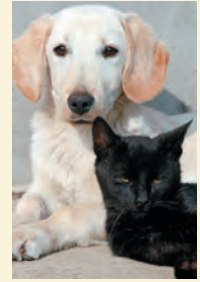
FOR EXAMPLE

Blocking

Recap: In 2007, pet food contamination put cats at risk, as well as dogs. Our experiment should probably test the safety of the new food on both animals.

Questions: Why shouldn't we randomly assign a mix of cats and dogs to the two treatment groups? What would you recommend instead?

Dogs and cats might respond differently to the foods, and that variability could obscure my results. Blocking by species can remove that superfluous variation. I'd randomize cats to the two treatments (test food and safe food) separately from the dogs. I'd measure their responses separately and look at the results afterward.



JUST CHECKING

2. Recall the experiment about gastric freezing, an old method for treating peptic ulcers that you read about in the first Just Checking. Doctors would insert an inflatable bladder down the patient's esophagus and into the stomach and then pump in a cold liquid. A major insurance company now states that it doesn't cover this treatment because "double-blind, controlled clinical trials" failed to demonstrate that gastric freezing was effective.
 - a) What does it mean that the experiment was double-blind?
 - b) Why would you recommend a placebo control?
 - c) Suppose that researchers suspected that the effectiveness of the gastric freezing treatment might depend on whether a patient had recently developed the peptic ulcer or had been suffering from the condition for a long time. How might the researchers have designed the experiment?

Adding More Factors

There are two kinds of gardeners. Some water frequently, making sure that the plants are never dry. Others let Mother Nature take her course and leave the watering to her. The makers of OptiGro want to ensure that their product will work under a wide variety of watering conditions. Maybe we should include the amount of watering as part of our experiment. Can we study a second factor at the same time and still learn as much about fertilizer?

We now have two factors (fertilizer at three levels and irrigation at two levels). We combine them in all possible ways to yield six treatments:

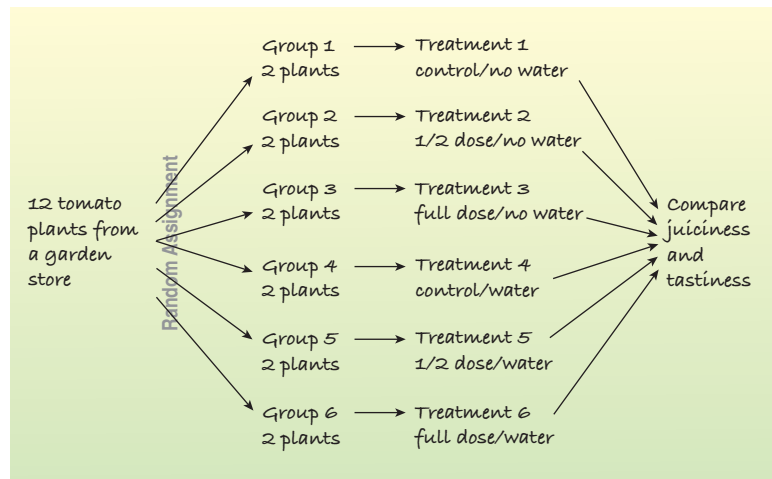
	No Fertilizer	Half Fertilizer	Full Fertilizer
No Added Water	1	2	3
Daily Watering	4	5	6

If we allocate the original 12 plants, the experiment now assigns 2 plants to each of these six treatments at random. This experiment is a **completely randomized two-factor experiment** because any plant could end up assigned at random to any of the six treatments (and we have two factors).

It's often important to include several factors in the same experiment in order to see what happens when the factor levels are applied in different *combinations*. A common misconception is that applying several factors at once makes it difficult to separate the effects of the individual factors. You may hear people say that experiments should always be run "one factor at a time." In fact, just the opposite

Think Like a Statistician

With two factors, we can account for more of the variation. That lets us see the underlying patterns more clearly.



is true: Experiments with more than one factor are both more efficient and provide more information than one-at-a-time experiments. There are many ways to design efficient multifactor experiments. You can take a whole course on the design and analysis of such experiments.

Confounding

Professor Stephen Ceci of Cornell University performed an experiment to investigate the effect of a teacher's classroom style on student evaluations. He taught a class in developmental psychology during two successive terms to a total of 472 students in two very similar classes. He kept everything about his teaching identical (same text, same syllabus, same office hours, etc.) and modified only his style in class. During the fall term, he maintained a subdued demeanor. During the spring term, he used expansive gestures and lectured with more enthusiasm, varying his vocal pitch and using more hand gestures. He administered a standard student evaluation form at the end of each term.

The students in the fall term class rated him only an average teacher. Those in the spring term class rated him an excellent teacher, praising his knowledge and accessibility, and even the quality of the textbook. On the question "How much did you learn in the course?" the average response changed from 2.93 to 4.05 on a 5-point scale.⁵

How much of the difference he observed was due to his difference in manner, and how much might have been due to the season of the year? Fall term in Ithaca, NY (home of Cornell University), starts out colorful and pleasantly warm but ends cold and bleak. Spring term starts out bitter and snowy and ends with blooming flowers and singing birds. Might students' overall happiness have been affected by the season and reflected in their evaluations?

Unfortunately, there's no way to tell. Nothing in the data enables us to tease apart these two effects, because all the students who experienced the subdued manner did so during the fall term and all who experienced the expansive manner did so during the spring. When the levels of one factor are associated with the levels of another factor, we say that these two factors are **confounded**.

In some experiments, such as this one, it's just not possible to avoid some confounding. Professor Ceci could have randomly assigned students to one of two classes during the same term, but then we might question whether mornings or

⁵ But the two classes performed almost identically well on the final exam.

afternoons were better, or whether he really delivered the same class the second time (after practicing on the first class). Or he could have had another professor deliver the second class, but that would have raised more serious issues about differences in the two professors and concern over more serious confounding.

FOR EXAMPLE

Confounding

Recap: After many dogs and cats suffered health problems caused by contaminated foods, we're trying to find out whether a newly formulated pet food is safe. Our experiment will feed some animals the new food and others a food known to be safe, and a veterinarian will check the response.

Question: Why would it be a bad design to feed the test food to some dogs and the safe food to cats?

This would create confounding. We would not be able to tell whether any differences in animals' health were attributable to the food they had eaten or to differences in how the two species responded.



A two-factor example Confounding can also arise from a badly designed multifactor experiment. Here's a classic. A credit card bank wanted to test the sensitivity of the market to two factors: the annual fee charged for a card and the annual percentage rate charged. Not wanting to scrimp on sample size, the bank selected 100,000 people at random from a mailing list. It sent out 50,000 offers with a low rate and no fee and 50,000 offers with a higher rate and a \$50 annual fee. Guess what happened? That's right—people preferred the low-rate, no-fee card. No surprise. In fact, they signed up for that card at over twice the rate as the other offer. And because of the large sample size, the bank was able to estimate the difference precisely. But the question the bank really wanted to answer was “how much of the change was due to the rate, and how much was due to the fee?” unfortunately, there's simply no way to separate out the two effects. If the bank had sent out all four possible different treatments—low rate with no fee, low rate with \$50 fee, high rate with no fee, and high rate with \$50 fee—each to 25,000 people, it could have learned about both factors and could have also seen what happens when the two factors occur in combination.

Lurking or Confounding?

Confounding may remind you of the problem of lurking variables we discussed back in Chapters 7 and 9. Confounding variables and lurking variables are alike in that they interfere with our ability to interpret our analyses simply. Each can mislead us, but there are important differences in both how and where the confusion may arise.

A lurking variable creates an association between two other variables that tempts us to think that one may cause the other. This can happen in a regression analysis or an observational study when a lurking variable influences both the explanatory and response variables. Recall that countries with more TV sets per capita tend to have longer life expectancies. We shouldn't conclude it's the TVs “causing” longer life. We suspect instead that a generally higher standard of living may mean that people can afford more TVs and get better health care, too. Our data revealed an association between TVs and life expectancy, but economic conditions were a likely lurking variable. A lurking variable, then, is usually thought of as a variable associated with both y and x that makes it appear that x may be causing y .

Confounding can arise in experiments when some other variable associated with a factor has an effect on the response variable. However, in a designed experiment, the experimenter *assigns* treatments (at random) to subjects rather than just observing them. A confounding variable can't be thought of as causing that assignment. Professor Ceci's choice of teaching styles was not caused by the weather, but because he used one style in the fall and the other in spring, he was unable to tell how much of his students' reactions were attributable to his teaching and how much to the weather. A confounding variable, then, is associated in a noncausal way with a factor and affects the response. Because of the confounding, we find that we can't tell whether any effect we see was caused by our factor or by the confounding variable—or even by both working together.

Both confounding and lurking variables are outside influences that make it harder to understand the relationship we are modeling. However, the nature of the causation is different in the two situations. In regression and observational studies, we can only observe associations between variables. Although we can't demonstrate a causal relationship, we often imagine whether x could cause y . We can be misled by a lurking variable that influences both. In a designed experiment, we often hope to show that the factor causes a response. Here we can be misled by a confounding variable that's associated with the factor and causes or contributes to the differences we observe in the response.

It's worth noting that the role of blinding in an experiment is to combat a possible source of confounding. There's a risk that knowledge about the treatments could lead the subjects or those interacting with them to behave differently or could influence judgments made by the people evaluating the responses. That means we won't know whether the treatments really do produce different results or if we're being fooled by these confounding influences.


WHAT CAN GO WRONG?

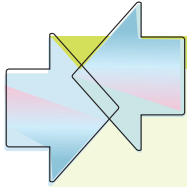
- ▶ **Don't give up just because you can't run an experiment.** Sometimes we can't run an experiment because we can't identify or control the factors. Sometimes it would simply be unethical to run the experiment. (Consider randomly assigning students to take—and be graded in—a Statistics course deliberately taught to be boring and difficult or one that had an unlimited budget to use multimedia, real-world examples, and field trips to make the subject more interesting.) If we can't perform an experiment, often an observational study is a good choice.
- ▶ **Beware of confounding.** Use randomization whenever possible to ensure that the factors not in your experiment are not confounded with your treatment levels. Be alert to confounding that cannot be avoided, and report it along with your results.
- ▶ **Bad things can happen even to good experiments.** Protect yourself by recording additional information. An experiment in which the air conditioning failed for 2 weeks, affecting the results, was saved by recording the temperature (although that was not originally one of the factors) and estimating the effect the higher temperature had on the response.⁶

It's generally good practice to collect as much information as possible about your experimental units and the circumstances of the experiment. For example, in the tomato experiment, it would be wise to record details of the weather (temperature, rainfall, sunlight) that might affect the plants and any facts available about their

⁶ R. D. DeVeaux and M. Szelewski, "Optimizing Automatic Splitless Injection Parameters for Gas Chromatographic Environmental Analysis." *Journal of Chromatographic Science* 27, no. 9 (1989): 513–518.

growing situation. (Is one side of the field in shade sooner than the other as the day proceeds? Is one area lower and a bit wetter?) Sometimes we can use this extra information during the analysis to reduce biases.

- ▶ **Don't spend your entire budget on the first run.** Just as it's a good idea to pretest a survey, it's always wise to try a small pilot experiment before running the full-scale experiment. You may learn, for example, how to choose factor levels more effectively, about effects you forgot to control, and about unanticipated confoundings. 



CONNECTIONS

The fundamental role of randomization in experiments clearly points back to our discussions of randomization, to our experiments with simulations, and to our use of randomization in sampling. The similarities and differences between experiments and samples are important to keep in mind and can make each concept clearer.

If you think that blocking in an experiment resembles stratifying in a sample, you're quite right. Both are ways of removing variation we can identify to help us see past the variation in the data.

Experiments compare groups of subjects that have been treated differently. Graphics such as boxplots that help us compare groups are closely related to these ideas. Think about what we look for in a boxplot to tell whether two groups look really different, and you'll be thinking about the same issues as experiment designers.

Generally, we're going to consider how different the mean responses are for different treatment groups. And we're going to judge whether those differences are large by using standard deviations as rulers. (That's why we needed to replicate results for each treatment; we need to be able to estimate those standard deviations.) The discussion in Chapter 6 introduced this fundamental statistical thought, and it's going to keep coming back over and over again. Statistics is about variation.

We'll see a number of ways to analyze results from experiments in subsequent chapters.



WHAT HAVE WE LEARNED?

We've learned to recognize sample surveys, observational studies, and randomized comparative experiments. We know that these methods collect data in different ways and lead us to different conclusions.

We've learned to identify retrospective and prospective observational studies and understand the advantages and disadvantages of each.

We've learned that only well-designed experiments can allow us to reach cause-and-effect conclusions. We manipulate levels of treatments to see if the factor we have identified produces changes in our response variable.

We've learned the principles of experimental design:

- ▶ We want to be sure that variation in the response variable can be attributed to our factor, so we identify and control as many other sources of variability as possible.
- ▶ Because there are many possible sources of variability that we cannot identify, we try to equalize those by randomly assigning experimental units to treatments.
- ▶ We replicate the experiment on as many subjects as possible.
- ▶ We consider blocking to reduce variability from sources we recognize but cannot control.

We've learned the value of having a control group and of using blinding and placebo controls.

Finally, we've learned to recognize the problems posed by confounding variables in experiments and lurking variables in observational studies.

Terms

Observational study	292. A study based on data in which no manipulation of factors has been employed.
Retrospective study	292. An observational study in which subjects are selected and then their previous conditions or behaviors are determined. Retrospective studies need not be based on random samples and they usually focus on estimating differences between groups or associations between variables.
Prospective study	293. An observational study in which subjects are followed to observe future outcomes. Because no treatments are deliberately applied, a prospective study is not an experiment. Nevertheless, prospective studies typically focus on estimating differences among groups that might appear as the groups are followed during the course of the study.
Experiment	294. An experiment <i>manipulates</i> factor levels to create treatments, <i>randomly assigns</i> subjects to these treatment levels, and then <i>compares</i> the responses of the subject groups across treatment levels.
Random assignment	294. To be valid, an experiment must assign experimental units to treatment groups at random. This is called random assignment.
Factor	294. A variable whose levels are manipulated by the experimenter. Experiments attempt to discover the effects that differences in factor levels may have on the responses of the experimental units.
Response	294. A variable whose values are compared across different treatments. In a randomized experiment, large response differences can be attributed to the effect of differences in treatment level.
Experimental units	294. Individuals on whom an experiment is performed. Usually called subjects or participants when they are human.
Level	294. The specific values that the experimenter chooses for a factor are called the levels of the factor.
Treatment	294. The process, intervention, or other controlled circumstance applied to randomly assigned experimental units. Treatments are the different levels of a single factor or are made up of combinations of levels of two or more factors.
Principles of experimental design	<ul style="list-style-type: none"> ▶ 295. Control aspects of the experiment that we know may have an effect on the response, but that are not the factors being studied. ▶ 296. Randomize subjects to treatments to even out effects that we cannot control. ▶ 296. Replicate over as many subjects as possible. Results for a single subject are just anecdotes. If, as often happens, the subjects of the experiment are not a representative sample from the population of interest, replicate the entire study with a different group of subjects, preferably from a different part of the population. ▶ 296. Block to reduce the effects of identifiable attributes of the subjects that cannot be controlled.
Statistically significant	299. When an observed difference is too large for us to believe that it is likely to have occurred naturally, we consider the difference to be statistically significant. Subsequent chapters will show specific calculations and give rules, but the principle remains the same.
Control group	301. The experimental units assigned to a baseline treatment level, typically either the default treatment, which is well understood, or a null, placebo treatment. Their responses provide a basis for comparison.
Blinding	301. Any individual associated with an experiment who is not aware of how subjects have been allocated to treatment groups is said to be blinded.
Single-blind	302. There are two main classes of individuals who can affect the outcome of an experiment:
Double-blind	<ul style="list-style-type: none"> ▶ those who could <i>influence the results</i> (the subjects, treatment administrators, or technicians). ▶ those who <i>evaluate the results</i> (judges, treating physicians, etc.). <p>When every individual in <i>either</i> of these classes is blinded, an experiment is said to be single-blind. When everyone in <i>both</i> classes is blinded, we call the experiment double-blind.</p>
Placebo	303. A treatment known to have no effect, administered so that all groups experience the same conditions. Many subjects respond to such a treatment (a response known as a placebo effect). Only by comparing with a placebo can we be sure that the observed effect of a treatment is not due simply to the placebo effect.
Placebo effect	303. The tendency of many human subjects (often 20% or more of experiment subjects) to show a response even when administered a placebo.

Blocking 303. When groups of experimental units are similar, it is often a good idea to gather them together into blocks. By blocking, we isolate the variability attributable to the differences between the blocks so that we can see the differences caused by the treatments more clearly.

Matching 304. In a retrospective or prospective study, subjects who are similar in ways not under study may be matched and then compared with each other on the variables of interest. Matching, like blocking, reduces unwanted variation.

Designs 298, 305. In a **completely randomized design**, all experimental units have an equal chance of receiving any treatment.

304. In a **randomized block design**, the randomization occurs only within blocks.

Confounding 306. When the levels of one factor are associated with the levels of another factor in such a way that their effects cannot be separated, we say that these two factors are confounded.

Skills

THINK

- ▶ Recognize when an observational study would be appropriate.
- ▶ Be able to identify observational studies as retrospective or prospective, and understand the strengths and weaknesses of each method.
- ▶ Know the four basic principles of sound experimental design—control, randomize, replicate, and block—and be able to explain each.
- ▶ Be able to recognize the factors, the treatments, and the response variable in a description of a designed experiment.
- ▶ Understand the essential importance of randomization in assigning treatments to experimental units.
- ▶ Understand the importance of replication to move from anecdotes to general conclusions.
- ▶ Understand the value of blocking so that variability due to differences in attributes of the subjects can be removed.
- ▶ Understand the importance of a control group and the need for a placebo treatment in some studies.
- ▶ Understand the importance of blinding and double-blinding in studies on human subjects, and be able to identify blinding and the need for blinding in experiments.
- ▶ Understand the value of a placebo in experiments with human participants.

SHOW

- ▶ Be able to design a completely randomized experiment to test the effect of a single factor.
- ▶ Be able to design an experiment in which blocking is used to reduce variation.
- ▶ Know how to use graphical displays to compare responses for different treatment groups. Understand that you should *never* proceed with any other analysis of a designed experiment without first looking at boxplots or other graphical displays.

TELL

- ▶ Know how to report the results of an observational study. Identify the subjects, how the data were gathered, and any potential biases or flaws you may be aware of. Identify the factors known and those that might have been revealed by the study.
- ▶ Know how to compare the responses in different treatment groups to assess whether the differences are larger than could be reasonably expected from ordinary sampling variability.
- ▶ Know how to report the results of an experiment. Tell who the subjects are and how their assignment to treatments was determined. Report how and in what measurement units the response variable was measured.
- ▶ Understand that your description of an experiment should be sufficient for another researcher to replicate the study with the same methods.
- ▶ Be able to report on the statistical significance of the result in terms of whether the observed group-to-group differences are larger than could be expected from ordinary sampling variation.