CHAPTER 11

Randomized Experiments

What You'll Learn

- Randomizing treatment can yield unbiased causal estimates.
- All the tools of statistical inference and hypothesis testing work in experimental settings for adjudicating between genuine effects and noise.
- Even with a randomized experiment, numerous complications can arise and must be planned around.
- When experimental subjects fail to comply with their experimental assignment, it is important to make comparisons based on randomized assignment.
- Even when researchers can't implement their ideal experiment, sometimes they can find instances in which their treatment of interest was randomized for non-research purposes. Such "natural experiments" are often fruitful, fortuitous opportunities to answer important causal questions.

Introduction

We love regression, so we had a pretty good time with chapter 10. But we can see how the message that you can basically never get unbiased estimates of causal relationships just by controlling for confounders might be something of a downer. We are going to try to make it up to you in the next three chapters. We'll do so by showing you that there are better ways to learn about causality. Those ways are called *research designs*. Using research designs to learn about causality often involves quite a bit of cleverness and creativity. This makes research design one of the most fun topics you must master in order to think clearly with data.

In this chapter, we consider the research design called a randomized experiment. Randomized experiments are great because, if you can randomize treatment, there are no confounders. So you can eliminate bias from your estimates. An analogy to randomized experiments also helps to explain why the approaches we consider in the next two chapters are called research designs. When you run a randomized experiment, you literally get to *design* the way in which treatment is assigned.

In chapters 12 and 13, we will turn to other research designs that are a little less "designy." In particular, those research designs are ways of trying to learn about causality from data that you observe in the wild—that is, where the world, rather than an

experimenter, decided how treatment would be assigned. But before we go there, let's spend some time learning about randomized experiments and how they work.

Breastfeeding

At the time of this writing, it is a virtual article of faith in the developed world that babies should be breastfed. Consider, for example, this official statement from the World Health Organization: "Adults who were breastfed as babies often have lower blood pressure and lower cholesterol, as well as lower rates of overweight, obesity and type-2 diabetes. There is evidence that people who were breastfed perform better in intelligence tests." Similarly, in 2011, the Surgeon General of the United States issued a call to action to support breastfeeding, which she said is "one of the most highly effective preventive measures a mother can take to protect the health of her infant." The accompanying report claimed that breastfeeding prevents a host of childhood scourges, including ear infection, eczema, diarrhea, respiratory disease, asthma, obesity, type 2 diabetes, leukemia, and sudden infant death syndrome (SIDS). Indeed, an enormous scientific literature documents the positive correlation between breastfeeding and good health outcomes for children.

But before you jump to conclusions about the causal benefits of breastfeeding, consider this fact. In developing countries, breastfeeding seems to be correlated with worse, not better, health outcomes for children. In countries as diverse as Ghana, Kenya, Egypt, Brazil, Peru, Bolivia, and Thailand, breastfeeding has been found to be correlated with malnutrition and decreased height and weight.

What is going on? Is it possible that breastfeeding is good for kids in the industrialized world and bad for kids in the developing world? Maybe, but let's make sure we are thinking as clearly as possible. You have already learned that correlation does not necessarily imply causation. And in this case, the comparison of mothers who do and do not breastfeed is most likely not an apples-to-apples one.

First, think about the developing world, where breastfeeding is negatively correlated with children's physical well-being. One possibility is that breastfeeding causes these adverse outcomes. It's also possible that some confounding factor, like poverty, causes both breastfeeding and malnourishment. Breastfeeding costs a lot in terms of a mother's time, but it doesn't cost much money. Formula, by contrast, costs a lot of money but less time. So we might expect that economically distressed families are more likely to breastfeed their children. And children from those same economically distressed families may be more prone to health problems for reasons entirely unrelated to breastfeeding. Reverse causality is also a concern. Perhaps adverse health in an infant directly leads a mother to be more likely to breastfeed.

Indeed, confounding factors and reverse causality appear highly relevant in the developing world. A 1997 study published in the *International Journal of Epidemiology* tracked the health outcomes of 238 toddlers in a village in Peru. The study's data included information on child size, breastfeeding, complementary food intake, and diarrhea. The study found a negative correlation between breastfeeding and size—children who were breastfed were smaller on average, suggesting they were in poorer health. This relationship was strongest among those children who were getting the least complementary food and were the most sickly. It turns out that, because breastfeeding is widely believed to have health benefits, mothers whose children were sickly or lacked access to complementary food weaned their children later. Consequently, children who

were already sick and malnourished were more likely to be breastfed. Thus, the study concludes, it is not breastfeeding that causes children not to grow in the developing world. Rather, children who are not growing because they are sick and malnourished are more likely to be breastfed.

Now think about the developed world, where parents are inundated with the message that breastfeeding is good for their children. Remember, breastfeeding is said to reduce the risk of heart disease, asthma, obesity, leukemia, SIDS, ear infections, and a host of other ailments. Unfortunately, the evidence underlying this conventional wisdom once again doesn't withstand much scrutiny. Surely you can think of lots of reasons why comparing breastfed children to children who are not breastfed might not be an apples-to-apples comparison. For instance, once official organizations issue statements on the efficacy of breastfeeding, we would expect wealthy, educated mothers to be particularly likely to hear this news and follow the advice. But their children were likely to have better health outcomes anyway.

Breastfeeding or not is such a high-stakes decision, and so many different factors influence this decision, that it might be impossible to find an apples-to-apples comparison out there in the world. However, perhaps we could generate our own apples-to-apples comparison through a randomized experiment. A team of researchers in Belarus tried to do exactly this.

The team's strategy was to run a randomized experiment. Clearly, for both ethical and practical reasons, they couldn't force mothers to breastfeed or not, just for the benefit of their study. But they could make it more likely that a randomly selected group of mothers would choose to breastfeed through randomly assigned encouragement. To achieve this, in some randomly selected hospitals the researchers implemented a program to encourage and facilitate breastfeeding. In other randomly selected hospitals, they did not implement this program. For all the hospitals, they recorded how children were fed and tracked a variety of the children's health (and other) outcomes over time. And, indeed, mothers in the hospitals that had the breastfeeding program were much more likely to breastfeed their newborns.

Despite the claims from the World Health Organization, the Surgeon General, the American Academy of Pediatrics, and the parenting industry more generally, they found surprisingly scant evidence for the large benefits breastfeeding is supposed to provide. Babies from the hospitals that received the program were slightly less likely to have eczema and gastrointestinal infections, but the researchers obtained null results (i.e., no statistically significant evidence of effects of breastfeeding) for many more outcomes. In a follow-up study, conducted when the children were between six and seven years old, the investigators explored whether the children whose mothers were encouraged to breastfeed performed better on any observable physical, psychological, or cognitive outcomes. They found no evidence that breastfeeding provided benefits in terms of the risk of eczema, allergies, asthma, obesity, emotional problems, conduct problems, hyperactivity, or peer problems. Indeed, if anything, the evidence went the other way, showing some limited evidence of a negative association between breastfeeding and these outcomes. The one piece of evidence they found in support of breastfeeding was that children from the hospitals that received the breastfeeding program performed slightly better on IQ tests. But in thinking about this one finding, don't forget the lessons about over-comparing we discussed in chapter 7. If you look at that many outcomes, you're pretty likely to find at least one statistically significant finding just because of noise.

Overall, our view is that the experimental evidence is not nearly strong enough to encourage every mother around the world to breastfeed. Although there are strong correlations between breastfeeding and health outcomes in various settings, and some reasonable arguments about the biological mechanisms through which breastfeeding might work, the best available evidence suggests that the average effect of breastfeeding is likely small. Without the power of randomized experimentation, it would be easy to over- or under-estimate the benefits of breastfeeding.

Randomization and Causal Inference

What makes randomized experiments such a powerful tool for learning about causal relationships? To start to see the answer, let's return to our discussion of potential outcomes and Body Vibes.

Suppose we want to know the effect of some treatment, say Body Vibes, on some outcome of interest, say skin health. In general, it is difficult to estimate the effect of Body Vibes because of all the issues discussed in previous chapters. We want to know how different a person's skin would be in the world in which they use Body Vibes versus the world in which they do not use Body Vibes. Unfortunately, for any given person, we only get to observe one of those potential outcomes. For example, if a person uses Body Vibes, we can observe their skin health in that situation, but we don't know what their skin would be like if they hadn't used Body Vibes.

If we just compared the average skin health of people who do and don't use Body Vibes, that's not comparing apples to apples. That is, there are a variety of confounders that imply that this difference in means is not an unbiased estimate of the average treatment effect. For instance, perhaps those who use Body Vibes just care more about their skin, and they also use more moisturizer and sunscreen. Or maybe the bias goes the other way. Perhaps the people who use Body Vibes have bad skin, they've tried everything else, and they're getting desperate. Either way, because of such confounders, we can't get an unbiased estimate of the effect of Body Vibes just by comparing the average skin health of people who do and do not use them.

One way, perhaps the best way, to get rid of this bias and be sure that we're making an apples-to-apples comparison is to randomize the treatment. Our comparison of those using and not using Body Vibes is biased because these groups likely have baseline differences. That is, on average, they would likely have different skin health even if none (or all) of them used Body Vibes. However, if we randomly assign people to use or not use Body Vibes, then those two groups would, in expectation, be the same in terms of their pre-existing skin health and all other prior characteristics. That is, there would be no confounders. Why is this the case?

If a treatment of interest is determined by the flip of a coin, a random-number generator on your computer, or another random process, then the only thing that distinguishes people in the treated group and the untreated group is pure chance. There is no reason why people in the treated group should be systematically taller, smarter, richer, more motivated, or have better skin than those in the untreated group before we deliver the treatment.

In terms of our potential outcomes notation, suppose we randomly assign some group (the treated group $\mathcal{T}$) to receive Body Vibes and another group (the untreated group $\mathcal{U}$) not to receive them. We observe the average skin health of the treated group with Body Vibes, $\overline{Y}_{1\mathcal{T}}$. And we observe the average skin health of the untreated group

without Body Vibes, $\overline{Y}_{0\mathcal{U}}$. Thus, if we compare the average skin health in the two groups we get the difference in means:

$$\overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}$$

But, because of randomization, there are no systematic differences between either of these groups and the population as a whole. Thus, the average skin health of the treated group wearing Body Vibes is an unbiased estimate of the average skin health of the whole population in the hypothetical world in which everyone wears Body Vibes:

$$\overline{Y}_{1\mathcal{T}} = \overline{Y}_1 + \text{Noise}_1$$

And similarly for the untreated group:

$$\overline{Y}_{0\mathcal{U}} = \overline{Y}_0 + \text{Noise}_0$$

Hence, the observed difference in means is an unbiased estimate of the average treatment effect,

$$\overbrace{\overline{Y}_{1\mathcal{T}} - \overline{Y}_{0\mathcal{U}}}^{\text{Observed Difference in Means}} = \overbrace{\overline{Y}_1 - \overline{Y}_0}^{\text{ATE}} + \text{Noise},$$

where this noise is just the difference of the two noise terms above.

As in the examples from previous chapters, noise can come from sampling variability. Perhaps there's a broader population about which we care, and we happened by chance to get an unusual sample of subjects in our experiment. It can also come from measurement error. And now, when we're doing an experiment, noise also comes from the random assignment of the treatment to subjects. Even for the same sample of subjects, different randomizations could have produced different estimates, and this also contributes to the noise.

Because of noise, for any small-scale experiment, there will be some differences in average potential outcomes between the treated and untreated groups, just by chance. But those differences won't be systematic—if we were to run many iterations of the experiment, we would not expect to find the same pattern of differences repeated over and over. This is what we mean by the phrase *in expectation* several paragraphs above.

Think back to our favorite equation:

$$\text{Estimate} = \text{Estimand} + \text{Bias} + \text{Noise}$$

Randomization guarantees that the bias is zero. So noise is the only reason that the estimate we get from comparing the mean outcome in the treated and untreated groups in a properly randomized experiment differs from the true causal effect (i.e., the estimand).

As we increase the number of subjects in any given experiment, we expect the two groups to become more and more similar. That is, as the sample size gets big, the noise becomes small.

Randomization gives you comparability of the treated and untreated group in expectation—generating unbiased estimates. A large sample size gives you a very small

amount of noise—generating precise estimates. So randomization plus a large sample size gives you comparability in your actual, realized sample—generating estimates that are very likely to be close to the true estimand.

If you think clearly about it, you'll realize that randomization is essentially the only way to guarantee unbiased estimates of causal relationships. Suppose you tried to conduct an experiment, but instead of randomly assigning subjects to the treated and untreated groups, you tried to carefully divide the groups so that they were as similar as possible to each other. Since you can't possibly observe and quantify all of your subjects' relevant characteristics, you'd have to make some judgment calls. Maybe you'll do this really well, and maybe you won't. What if your own subconscious biases lead you to put slightly different people in the treated and untreated groups—perhaps because you're subconsciously hoping the experiment will show a big effect? You'll have no way of knowing whether you actually did a good job. Therefore, why take the risk? Why not actually flip a coin and assign the treatment randomly? If this point seems obvious now, it wasn't obvious to a lot of smart people in the past. It's only since the work of R. A. Fisher in the 1920s that scientific researchers have understood the value of randomization.

There is one way in which this thought about trying to make the treated and untreated groups as similar as possible on observable characteristics does make some sense. As we know from our favorite equation, there are two ways our estimates might differ from the true causal effect: bias and noise. Randomization eliminates bias. But there could still be lots of noise, especially if the sample size is small or the experimental subjects are quite different from one another on characteristics that matter for the outcome. That is, in any given iteration of an experiment, the treated and untreated groups could end up looking very different *in reality*, even though they are the same *in expectation*.

One thing you can do to reduce this problem is to start by grouping people on the basis of their observable similarities. Then you can randomly assign individuals to be treated or untreated within those groupings. This is called *blocking* or *stratification*. For example, you might be concerned that men and women, on average, have very different levels of skin health. It would introduce a lot of noise into your experiment if, by chance, you ended up with a treated group made up of mostly men and an untreated group made up of mostly women, or vice versa. This won't happen in expectation (i.e., if you did the experiment an infinite number of times, the average proportion of men and women will be the same in the treated and untreated groups). But it could happen in any given iteration of your randomization. To eliminate this source of noise, you could start by dividing your experimental population by biological sex. Then you randomly assign half the male group to be treated and half to be untreated, and likewise for the female group. You'd still have randomized treatment assignment, so you'd still get unbiased estimates. But you'd also have reduced the noise by making sure that your treated and untreated groups were similar in terms of biological sex, not just in expectation but in reality.

Extending this logic, an analyst could identify many different blocks or strata of subjects with similar pre-treatment characteristics and conduct their randomization within these strata or blocks. The most extreme version of this would be a matched-pair design, where an analyst identifies pairs of individuals that they believe are most similar to one another and for each pair randomly assigns one to be treated and the other to be untreated. This can be a great way to improve the precision of one's estimates. But you must make sure that the treatment is assigned randomly within each pair.

Estimation and Inference in Experiments

In chapter 6, we discussed statistical inferences about relationships. All of those lessons apply in the case of an experimental estimate as well. In the simplest scenario, we can analyze the results of an experiment by calculating a difference in means—that is, comparing the average outcome in the treated and untreated groups. In fact, as we saw in chapter 5, if we regress the outcome on a binary measure of treatment status, the regression coefficient associated with the treatment variable is just the difference in means. And since regression coefficients and differences in means are just quantitative relationships, we can apply all of the statistical tools of chapter 6 to experiments as well.

Standard Errors

Suppose we conduct a randomized experiment and estimate the average treatment effect by comparing the average outcome for subjects with the treatment to the average outcome for subjects without the treatment. This estimate is unbiased. But it might be imprecise (i.e., there may be lots of noise).

We'd, of course, like to know how close our estimates are to the true effect of interest (i.e., the estimand). We can estimate the standard error associated with our experimental estimate just like we estimated the standard error of poll results and regression coefficients in chapter 6. The standard error gives us a sense of how far, on average, our estimate would be from the truth as a result of noise if we repeated our experiment an infinite number of times, each time using the same procedure to generate an estimate of the treatment effect. Similar to our discussion of poll results, the true standard error depends on quantities that are unobservable, but there are various approximations that practitioners use for estimating the standard error.

You don't need to memorize formulas for standard errors; you can always look them up or just let your computer calculate them for you. Nonetheless, it's useful to think about how various features of experiments influence the amount of noise. Suppose we conduct an experiment with N subjects, of whom m receive the treatment and $N - m$ do not. All else equal, the greater N is, the less noise and, thus, the smaller the standard error. This should be intuitive. When the sample size is larger, the treated and untreated groups will be more similar to each other with respect to other characteristics, reducing noise, and making our estimates closer to the true causal effect.

What about m? Suppose we have five hundred people in our study. How many of them would we like to put in the treated group and in the untreated group? Obviously, we can't put all of them in either group because then we wouldn't be able to make a comparison (remember that correlation requires variation). Extending that logic, we can see that we don't want too few subjects in either group. If either the treated group or the untreated group is very small, then our estimates will be imprecise because the average outcome for whichever group is small will be quite sensitive to the idiosyncratic features of just a few subjects. Typically, then, you'll get the most precise estimates when the sizes of the treated and untreated groups are roughly equal.

With that being said, there are often cases where an optimal experimental design might have different numbers of subjects in each condition. Suppose you have 100,000 potential subjects. You only have enough resources to put 100 people in your treated group, but it's costless to put more subjects in your untreated group. You might as well randomly assign 100 people to the treated group and everybody else to the

untreated group. Your estimates won't be nearly as precise as if you had 50,000 people in each group, but they'll be much more precise than an experiment with 100 people in each group.

The last factor that influences the noisiness of experimental estimates is how variable the outcome is in both the treated and untreated groups. If we study outcomes with little variance within each treatment condition, our estimates will be more precise than if we study outcomes with greater variance. This is because, if the outcome doesn't vary much based on non-treatment characteristics, then there is very little scope for noise—we'll get similar outcomes for each group across iterations of the experiment. This is why, for example, doctors and government regulators often have precise estimates of the effect of some heart medication on blood pressure (a relatively low variance outcome) but imprecise estimates of the effect of the same drug on heart attacks (a high variance outcome). Of course, sometimes, we have no control over this. The outcome of interest is what it is. And sometimes the most interesting or important outcomes (e.g., heart attacks) are high variance. But other times it might be possible to identify outcomes or methods for measuring those outcomes that reduce noise.

Hypothesis Testing

We can also apply the tools of hypothesis testing that we learned in chapter 6 to experimental results to assess statistical significance. For instance, dividing the estimate by the standard error generates a value called a t-statistic, which can be used to estimate a p-value. And because we often do hypothesis testing with experimental results, we need to keep thinking clearly about the risks of over-comparing, under-reporting, and reversion to the mean. Recall from chapter 7 that analysts can reduce these risks by stating up front the questions of interest the experiment is designed to address, pre-specifying the hypotheses they plan to test and regressions they plan to run (so they can't just go fishing for a statistically significant finding), and reporting the results regardless of what they find.

We should also interpret experiments with these issues in mind. If analysts are not transparent about the steps they took to avoid over-comparing and under-reporting, we should be skeptical of their findings. And, the more surprising the findings, the more skeptical we should be. Remember that the ESP result arose in an experimental study! Or, more seriously, think again about the breastfeeding experiment with which we began this chapter. That study had many virtues. But one potential problem is that, because the study designers collected information on so many outcomes, when we see no evidence of an effect of breastfeeding on eczema, allergies, asthma, obesity, emotional problems, conduct problems, hyperactivity, or peer problems, but we do see an effect on IQ, we are worried that the apparent effect on IQ arose just by chance.

Problems That Can Arise with Experiments

Things rarely work as beautifully in practice as they do in our idealized examples. In theory, you can design a randomized experiment and estimate the average treatment effect simply by comparing means. In practice, however, problems arise that make analysis and interpretation less straightforward. Let's discuss some of those problems and the ways in which careful analysts can deal with them. Thinking about these issues now will have benefits beyond the context of experiments because these problems can arise for virtually any strategy for estimating causal relationships.

Noncompliance and Instrumental Variables

One common problem in experiments is that subjects fail to comply with their assigned treatment. We call this *noncompliance*. For instance, it is pretty common in medical studies for some subjects to simply stop taking their medication. There was also noncompliance in the breastfeeding experiment. Recall, because it is unethical to force a mother to breastfeed or not, the researchers randomly assigned mothers into groups where they received more or less encouragement to breastfeed. Encouragement designs like this allow researchers to experimentally study lots of topics that would otherwise be off-limits for logistical or ethical reasons. But such studies inevitably involve the additional complications that arise from noncompliance, since surely some mothers who were encouraged nonetheless did not breastfeed and some mothers who were not encouraged did breastfeed.

Suppose we designed a randomized experiment to estimate the effect of Body Vibes on skin health. We randomly assign some individuals to the treatment condition—we give them Body Vibes and, for the sake of science, try to convince them to wear them. We also randomly assign some individuals to an untreated condition—they are given no Body Vibes and told to go about their normal lives. Then, despite our best efforts, some of the subjects in the treated group forget or simply refuse to wear their Body Vibes. And amazingly, a few of the more gullible members of the untreated group hear about Body Vibes elsewhere, spend their hard-earned money on the product, and wear them. Shoot! What do we do?

One idea would be to simply compare people who did and didn't wear Body Vibes, ignoring whether each subject was initially assigned to the treated or untreated group. But this won't work. It brings us right back to the problem we were trying to solve through our randomized experiment. The people who voluntarily wear or do not wear Body Vibes are likely different from one another, so a comparison of those two groups is not apples-to-apples.

Another idea would be to drop the subjects that did not comply with their treatment assignments. In other words, we could remove from our analysis the people who were assigned to the treated group but didn't wear Body Vibes and the people who were assigned to the untreated group but did wear them. After doing that, we might just proceed as normal, comparing the mean skin health among the remaining members of the treated and untreated groups.

Unfortunately, this is still not an adequate solution. To see why, think about the people who were in the treated group but refused to wear Body Vibes. They might be special in important ways—for example, they may have better skin health or be less gullible. Presumably, there were also people just like them in the untreated group. But, because we didn't ask those people to wear Body Vibes in the first place, we can't figure out who they are. So we can't similarly remove them from the untreated group. Thus, if we throw this group of people out of the treated group, the comparison of the treated and untreated groups will no longer be apples-to-apples. The kinds of people who wouldn't wear Body Vibes even if given to them would be present in the untreated group but not the treated group.

So what can we do in light of noncompliance? Well, one thing we can always do is estimate the effect of being assigned to the treated group (as distinct from the effect of the treatment itself). We sometimes call this the *intent-to-treat* (ITT) effect or the *reduced-form* effect. We do this by comparing the outcomes for the people assigned to the treated and untreated groups, regardless of whether they actually comply with their

treatment assignment. This comparison won't give us an unbiased estimate of the effect of wearing Body Vibes. But it will give us an unbiased estimate of the effect of being given Body Vibes and encouraged to wear them.

There are situations where a policy maker or decision maker actually cares more about intent-to-treat effects than actual treatment effects. Suppose a charitable organization is trying to decide whether it should provide free Body Vibes to high school kids with bad skin. They know that not everyone provided with Body Vibes will wear them. Furthermore, all they can do from a policy perspective is provide the Body Vibes; they can't force anyone to use them. They conduct an experiment to estimate the benefits of free Body Vibes. What quantity should go on the benefits side of their cost-benefit analysis to inform them about whether this is a good policy? It's not the average effect of Body Vibes for an individual who uses them. It's the average effect of being provided Body Vibes, regardless of whether an individual uses them or not, since this is what the charitable organization can actually do. So the intent-to-treat effect is the relevant number. More seriously, in many settings, all a policy maker or organization can do is provide a service; they can't force people to take it up. In any such situation, the intent-to-treat effect may in fact be the most important quantity.

In other situations, however, we are interested in the actual effect of the treatment, not just the intent-to-treat effect. Suppose, for instance, that we're trying to decide whether we should wear Body Vibes ourselves. Or, more seriously, suppose someone is deciding whether to try an experimental medical treatment, a new study regimen, a new teaching technique, or a new productivity-enhancing management strategy. In those cases, we want to know more than just the intent-to-treat effect. We want to know the likely effect of taking up the treatment. So what more can we do with our experimental results, plagued as they are by issues of noncompliance?

To make some additional progress, let's think about the different ways a subject can respond to our experimental encouragement to wear or not wear Body Vibes. Our sample consists of up to four different kinds of people.

1. There are *compliers*, who will wear Body Vibes if they're assigned to treatment and will not wear them if they're not assigned to treatment.
2. There are *always-takers*, who will wear Body Vibes regardless of whether or not they are assigned to treatment.
3. There are *never-takers*, who will not wear Body Vibes regardless of whether or not they are assigned to treatment. (We are both never-takers when it comes to Body Vibes.)
4. And, in principle, there could be a perverse group of *defiers*, who won't wear Body Vibes if they're in the treated group but will wear Body Vibes if they're in the untreated group.

Obviously, when we do an experiment, we're hoping for lots of compliers. The whole idea of an experiment is that we want to randomly assign treatment, and the compliers are those subjects who are willing to let us do that.

Every subject in an experiment fits neatly into one (and only one) of these categories. However, we can't just look at our experimental subjects and figure out which people are compliers, always-takers, never-takers, or defiers. Why is that? Suppose we see that someone is in the untreated group and doesn't wear Body Vibes. We know that they are either a complier or a never-taker. But we have no way of knowing which, because we don't know whether they would have worn Body Vibes if they were in

Table 11.1. Who takes the treatment in a Body Vibes experiment?

	Treated Group	**Untreated Group**
Wore Body Vibes	Compliers & Always-Takers	Always-Takers & Defiers
Didn't Wear Body Vibes	Never-Takers & Defiers	Compliers & Never-Takers

Table 11.2. Who takes the treatment in a Body Vibes experiment, assuming there are only compliers and never-takers?

	Treated Group	**Untreated Group**
Wore Body Vibes	Compliers	N/A
Didn't Wear Body Vibes	Never-Takers	Compliers & Never-Takers

the treated group. Table 11.1 illustrates this issue more generally for our Body Vibes experiment.

Dividing people up into these groups helps us think clearly about when we are or are not making an apples-to-apples comparison. In particular, in order to ensure that we don't have confounding, we want the groups we compare (say, treated and untreated groups) to have the same share of compliers, always-takers, never-takers, and defiers.

To get a sense of how this helps us understand the problem, let's start by assuming that everyone is either a complier or a never-taker. In other words, none of those people who might buy and wear Body Vibes on their own happened to participate in our experiment. (We'll relax this in a bit.) Table 11.2 shows what our experiment looks like in a world with only compliers and never-takers.

Now let's revisit the various ways we might deal with experimental subjects who don't behave according to their treatment assignment. It's easy to see why we can't just compare people who did and didn't wear Body Vibes, ignoring their treatment assignment. The group that wears Body Vibes is made up of just the compliers in the treated group. The group that doesn't wear Body Vibes is a combination of the compliers in the untreated group, the never-takers in the untreated group, and the never-takers in the treated group. So the comparison of Body Vibes wearers to non–Body Vibes wearers is not apples-to-apples.

Similarly, it's easy to see why we can't just drop the people who visibly don't comply with our experiment. We would drop the never-takers from the treated group. But we wouldn't drop anyone from the untreated group. As a result we'd be comparing the compliers from the treated group to a combination of the compliers and the never-takers from the untreated group—again, not an apples-to-apples comparison.

It seems like we're still stuck in a place where all we can do is compare the treated and untreated groups, estimating the intent-to-treat effect. But, actually, we can do better. Let's see how.

A key step in doing better involves estimating the proportion of compliers in our sample. We don't know exactly who the compliers are. But, in our simplified example with only compliers and never-takers, we can estimate what proportion of the sample is compliers. We do so by calculating the proportion of the treated group that takes up the treatment. This is the proportion of compliers in the treated group. And, because

Table 11.3. Observed differences between the two experimental groups.

	People Assigned to Be Treated	People Assigned to Be Untreated
Average Skin Health	7.8	6.2

of random assignment, the treated and untreated groups have the same proportion of compliers in expectation. Therefore, the proportion of compliers in the treated group is an unbiased estimate of the proportion of compliers in the whole sample (i.e., the treated group and untreated group combined).

We now have unbiased estimates of both the intent-to-treat effect (by comparing the average outcomes in the group assigned to be treated and the group assigned to be untreated) and the proportion of compliers in the sample. How does that help us?

We want to know the effect of Body Vibes on some outcome like skin health. If we assume that the only way that treatment assignment could have influenced skin health is through the actual use of Body Vibes, then what is the intent-to-treat effect? Under our assumption, the never-takers were not affected by the treatment assignment, and the effect of the treatment assignment for the compliers is just the effect of Body Vibes. So the expected intent-to-treat effect is the average effect of Body Vibes for compliers times the proportion of compliers in the sample. That means if we divide the ITT effect by our estimate for the proportion of compliers, we'll have an unbiased estimate of the average effect of the treatment for compliers.

Let's do a little example to see how this works. Imagine that Body Vibes actually work. (Remember, a lot of this book is about counterfactual worlds.) In particular, suppose that you could measure skin health on a scale of 1–10, with 10 being perfect skin and 1 being very bad skin.

Now let's imagine we conducted an experiment on 100 people to study the effects of Body Vibes. We randomly assigned 50 people to receive the treatment and 50 people not to. The people assigned to receive the treatment got Body Vibes. The other people did not. A month later, we measured the skin health of each person on our 1–10 scale. Suppose the data looked like that in table 11.3.

Our estimate from the data of the intent-to-treat effect is 1.6—that is, on average, people given Body Vibes had a skin health score that was 1.6 points higher than people not given Body Vibes.

You dig a little deeper and discover that, while no one in the group assigned to be untreated went and bought Body Vibes, only 40 of the 50 people assigned to treatment wore them. From this you estimate that the proportion of compliers in your sample is 80 percent ($\frac{40}{50}$) and the proportion of never-takers in your sample is 20 percent. You can now estimate the true effect of Body Vibes on the compliers.

How does this work? To make sure we are thinking clearly, let's return to our potential outcomes notation. Let Y_{0c} be the average skin health of a complier without treatment (i.e., without Body Vibes); let Y_{1c} be the average skin health of a complier with treatment (i.e., with Body Vibes); and let Y_{0n} be the average skin health of a never-taker without treatment. Given that we have 80 percent compliers and 20 percent never-takers, we have the following two equations:

$$7.8 = 80\% \cdot Y_{1c} + 20\% \cdot Y_{0n}$$

$$6.2 = 80\% \cdot Y_{0c} + 20\% \cdot Y_{0n}$$

The first equation says that the average skin health for those assigned to the treated group (7.8) is a weighted average of the average skin health of compliers with treatment (with weight 80%) and of never-takers without treatment (with weight 20%). Similarly, the average skin health for those assigned to the untreated group (6.2) is a weighted average of the average skin health of compliers without treatment (with weight 80%) and of never-takers without treatment (with weight 20%).

We can subtract the left-hand sides of these two equations from one another and the right-hand sides of these two equations from one another to get

$$1.6 = 80\% \cdot (Y_{1c} - Y_{0c}).$$

The left-hand side is the intent-to-treat effect: the difference in average outcomes between the group assigned to be treated and the group assigned to be untreated. On the right-hand side, "80%" represents the proportion of compliers in the sample. And the term in parentheses is the average effect of the treatment for compliers (usually called the *complier average treatment effect* or CATE). So, we can recover the complier average treatment effect by dividing both sides by 80 percent:

$$\begin{aligned} \frac{1.6}{80\%} &= \overbrace{Y_{1c} - Y_{0c}}^{\text{CATE}} \\ &= 2. \end{aligned}$$

It is important to note the distinction between the complier average treatment effect and the overall average treatment effect. It is possible that wearing Body Vibes has the same effect on skin health for everyone. In this scenario, we would say that there are *homogeneous treatment effects*. But this need not be the case—Body Vibes could differentially affect the skin health of different people, and the average effects might be quite different for the kind of person who would never use them (never-takers) and the kind of person who uses them if encouraged (compliers). In this case, we say there are *heterogeneous treatment effects*. As the algebra above shows, dividing the intent to treat effect by the share of compliers estimates the complier average treatment effect. If there are homogeneous treatment effects, the complier average treatment effect is the same as the overall average treatment effect. But if there are heterogeneous treatment effects, they are not the same and we have to keep in mind that we are only able to estimate the average treatment effect for this specific subgroup. The intuition for why is straightforward. It is only the compliers who are actually changing their behavior in response to treatment. So they are the only part of the population about whom we are actually gaining information.

It was relatively easy to see how all this works in a simplified world where everyone was either a complier or a never-taker. But we can do the same basic thing even if we move away from this simplified world and also allow for the possibility of always-takers. For now, let's continue to assume that there are no defiers, because they muddy the waters. (There are lots of situations, including this hypothetical Body Vibes experiment, where we think that there will be few to no defiers.)

Table 11.4 shows how different types of subjects appear in our experimental sample in this more complicated world.

How do we estimate the proportion of compliers when there are compliers, never-takers, and always-takers? First, the people in the group assigned to be treated who

Table 11.4. Who takes the treatment in a Body Vibes experiment, assuming there are no defiers?

	Treated Group	**Untreated Group**
Wore Body Vibes	Compliers & Always-Takers	Always-Takers
Didn't Wear Body Vibes	Never-Takers	Compliers & Never-Takers

actually wear Body Vibes are either compliers or always-takers. So the size of this group gives us an estimate of the proportion of always-takers plus compliers. Second, the people in the group assigned to be untreated who wear Body Vibes are definitely always-takers. So the size of this group gives us an estimate of the proportion of always-takers. By subtracting this second number from the first, we get an estimate of the proportion of compliers. With that in hand, we can again proceed as above—calculating the ITT effect and dividing it by the share of compliers to get the CATE.

Therefore, our general procedure for estimating the complier average treatment effect is as follows. First, estimate the ITT effect—that is, the effect of being assigned to the treated group on the outcome of interest. Second, estimate the effect of being assigned to the treated group on the actual take-up of the treatment. This is sometimes called the *first-stage effect*. Assuming there are no defiers, this gives us an unbiased estimate of the proportion of compliers. We then recover an estimate of the CATE by dividing the intent-to-treat effect by the proportion of compliers. This ratio is called the *Wald Estimator*, after the statistician Abraham Wald, who first developed it, though in a different context.

The Wald Estimator is a special case of what is called *instrumental variables* (IV) analysis. This kind of analysis is appropriate when the treatment of interest is not randomly assigned but there is some other variable (called an instrument) that (1) affects the treatment of interest, (2) does not affect the outcome of interest except through the treatment, and (3) is randomly assigned (or, there is some other way to credibly estimate its effect on the treatment and the outcome).

To be more precise, there are four key conditions that must hold for IV analysis to work:

1. **Exogeneity**: The instrument must be randomly assigned or be "as if" randomly assigned, allowing us to obtain unbiased estimates of both the first-stage and reduced-form (ITT) effects.
2. **Exclusion restriction**: All of the reduced-form effect must occur through the treatment. In other words, there is no other pathway for the instrument to influence the outcome except through its effect on the treatment. If this isn't the case, then the reduced-form effect includes both the effect of the treatment on the outcome for compliers and these other pathways. Then, even after we divide by the first-stage effect, the resulting estimate still includes these other pathways and, thus, does not reflect the CATE.
3. **Compliers**: There must be some compliers.
4. **No defiers**: If there are defiers, then our estimate will give us a weighted average of the average effect for compliers and the average effect for defiers, but with the defiers getting negative weight (since their behavior changed in the wrong direction). How big a problem the presence of defiers is depends on how many of them there are and how different the treatment effects are for compliers

and defiers. If there are very few defiers, then the bias that comes from their presence is negligible. But if there are many defiers, they are a big problem for the IV analysis.

In the case of our Body Vibes experiment, the experimental assignment to be treated or untreated was an excellent instrument. It clearly satisfied exogeneity because we randomized treatment. It also seems unlikely that being assigned to the treated group had any way of affecting skin health other than through Body Vibes, so it quite plausibly satisfied the exclusion restriction. So, as long as there were compliers (i.e., people who actually used the Body Vibes because they were assigned to) and no defiers, our analysis yielded an estimate of the complier average treatment effect.

There are more flexible ways to implement IV analysis than the Wald Estimator. In particular, it can be implemented using regression, which is important because that allows us to accommodate control variables, if necessary, as well as situations with multiple instruments or treatments and instruments that are not binary.

Some analysts think of IV as a method or research design unto itself. For example, an analyst might implement our design above and say that they estimated the effect of Body Vibes using instrumental variables. That's technically true but misleading. The important research design in our example is the randomized experiment. We're using instrumental variables to deal with noncompliance, acknowledging the additional assumptions (above and beyond randomization) that doing so requires. In particular, the exclusion restriction is defensible in our example because all the experiment did was hand out stupid stickers to some people and not to others. In other contexts, however, the exclusion restriction will be harder to justify and will require a lot of thought. We will return to this later, when we discuss natural experiments.

Chance Imbalance

Randomization guarantees that the treated and untreated groups are, in expectation, the same in terms of potential outcomes. But the term *in expectation* is important. Just because two groups are the same in expectation doesn't mean they are the same in actuality. As we've discussed, in any given experiment, the treated group could differ from the untreated group in lots of ways, just due to chance, and we might call this a *chance imbalance*. This is why there is a noise term, in addition to a bias term, in our favorite equation.

Experimenters often assess the *balance* between their treated and untreated groups by comparing them in terms of measurable pre-treatment characteristics. For example, in our Body Vibes experiment, we could compare the average age, gender, weight, diet, and skin health of the subjects in the treated and untreated groups before the treatments are delivered. We could even test for statistically significant differences. The hope, of course, is that we don't find any differences. If we do, we must worry that, even though our estimate is unbiased, it might nonetheless be quite different from the true effect because of noise.

What should a careful analyst do if, despite randomization, the treated and untreated groups turn out to differ in substantively or statistically significant ways? Let's consider three potential responses.

1. **Throw out the "broken" experiment.** You had good intentions when you ran the experiment, but you got unlucky and now you can't trust your results, so

you should just forget the experiment and move on. Maybe you should do another one and hope for better balance.

We think this is an inappropriate response. Remember the problem of overcomparing. If you test for balance on enough pre-treatment variables, you are virtually guaranteed to find statistically significant imbalance on some of them. Therefore, by this logic, the more pre-treatment variables you can measure, the more likely you are to have to throw out the experiment, which seems perverse. Moreover, even "broken" experiments contain information. Importantly, they are unbiased (remember, bias is about getting the answer right on average, across lots of iterations of the experiment). And, so, the information could be pooled with other evidence (perhaps from other iterations of the same experiment) and incorporated into a larger analysis that will ultimately contribute to knowledge.

Our response here assumes that the analyst is confident that the treatment was indeed randomly assigned. Our recommendations would change if this wasn't the case. Suppose you (or your computer) didn't do the randomization directly. Instead, suppose you were running a large-scale experiment and the randomization was implemented by a big team or by a partner organization. In a situation like this, if you detect enough imbalance, you might start to worry that your planned randomization wasn't faithfully implemented. In that case, throwing out the experiment (probably following some investigation into whether your suspicions are well founded) could be appropriate.

2. **Proceed as normal.** Unbiasedness is a property *in expectation*, so the experimental estimate is still unbiased. You could report the imbalance for the sake of transparency while still estimating the treatment effect as you originally planned. Of course, the treated and untreated groups are sometimes different by chance. That's exactly why we report standard errors or other measures of noise.

   This strategy may seem unsatisfying. As you'll recall from chapter 6 and our favorite equation, even an unbiased estimate can be very far from the truth. When we find an imbalance between the treated and untreated groups that we think is strongly related to the outcome of interest, we might worry that this chance imbalance reflects getting one of those draws of our procedure that result in an estimate that is far from the truth, despite the absence of bias. Nonetheless, there is still some merit to proceeding as planned and reporting your unbiased (if probably quite wrong) estimate. This is especially the case if we are talking about the kind of experiment that will be replicated lots of times, so that the lack of balance in any one iteration will be washed out in the long run through averaging across many iterations of the experiment.

   But we also might wonder whether there is some way that we can account for the imbalance and generate an estimate that is likely to be closer to the right answer right now—which leads us to our third possible response.

3. **Use the techniques discussed in chapter 10 to control for any unbalanced variables.** As we learned in chapter 10, controlling for pre-treatment variables could improve precision by accounting for the variance in the outcome that is due to those variables. This is the sense in which controlling may help you get closer to the truth. But it has disadvantages as well. Because of randomization, you can be sure that estimating the treatment effect without controlling (e.g., just comparing the average outcome in the treated and untreated groups) leads

to an unbiased (if potentially very far from correct) estimate of the true effect. By contrast, controlling for variables after the fact can produce a biased (if more precise) estimate. This means that if you were to run your experiment lots of times and always control for whatever variables turn out to be imbalanced, the average of your estimates might not converge on the true effect. So there are trade-offs to think about between reducing noise and increasing bias.

Another concern with this approach is that by controlling for pre-treatment variables, the researcher is exercising additional degrees of freedom that should raise concerns about over-comparing and under-reporting. As we learned in chapter 7, savvy consumers should be skeptical when they see an analyst play around with their specification, and if an experimental result depends upon a particular set of control variables that were not necessitated by the design, we probably shouldn't have much faith in that result.

There is no easy answer or quick fix to the problem posed by chance imbalance following randomization. Our view is that you should probably do some combination of options 2 and 3. Also, whenever feasible, you should try to replicate experiments multiple times. No matter what, be honest and transparent about the choices you make.

Of course, what we'd really like is to avoid these difficult decisions by avoiding chance imbalance in the first place. And there are ways to do this. If you can identify and measure important characteristics ahead of time, you can design your experiment to ensure balance. We've already briefly mentioned how—by using a *blocked* or *stratified* experimental design. Prior to treatment, divide your sample into groups based on those characteristics and then randomize within those groups. Recall that earlier in this chapter we suggested that you might be concerned that Body Vibes differentially affect men and women, so you want to make sure your treated and untreated groups are balanced by biological sex. You achieve this by first dividing your sample into a male group and a female group. Then you randomize treatment assignment within these groups. This guarantees that biological sex is balanced between the treated and untreated groups (reducing noise), while still assigning treatment randomly (preserving unbiasedness). We can save ourselves a lot of headaches by following a procedure like this for pre-treatment characteristics that would cause us concern if they turned out to be imbalanced after the fact.

Lack of Statistical Power

Sometimes, an otherwise excellent experiment yields inconclusive results because the standard error is so large that we don't learn much, and even a reasonably sized effect would not be statistically distinguishable from zero. In this case, we say that the experiment lacked the *statistical power* to detect the effect of interest. Ideally, an experimenter would think about this problem beforehand and take steps to improve the precision and statistical power of the experiment—for instance, by increasing the sample size.

That said, sometimes, because of costs or other constraints, it turns out that you've run an underpowered experiment. If you've already run the experiment and obtained imprecise estimates, what can you do? Here, the debate mirrors that around chance imbalance. You can try to improve precision by controlling for some variables, but, as we've already discussed, that has downsides. Sometimes, you may just have to accept that you don't have a convincing answer to your question and you haven't learned much, even after running an experiment.

Thinking back to chapter 7, you might be wondering whether burying the results of an underpowered experiment contributes to the file-drawer problem. The answer is, yes. And this is a good reason not to run underpowered experiments. But if the results of an experiment are so imprecise that we learn virtually nothing, there isn't much use publicizing them. So failing to publish because an experiment didn't teach us much is not nearly as detrimental to the scientific process as failing to publish because an experiment didn't give the desired result.

Attrition

Sometimes people drop out of an experiment after treatment assignment. Such *attrition* is importantly different from noncompliance. Noncompliance involves people who were supposed to take up the treatment but chose not to. At least we get to observe the outcome for these noncompliers. When people drop out of the experiment, we don't even get to observe their outcome.

Suppose, for example, that Body Vibes make some people feel so young and carefree that they forget to come back for their follow-up meeting where we were planning to measure their skin health. This is bad. If attrition happens at random (i.e., is unrelated to the treatment or the potential outcomes), then we can still obtain an unbiased estimate of the effect of our treatment by comparing the remaining members of the treated and untreated groups. We just lose some statistical power because our sample got smaller. If attrition is nonrandom but unaffected by the treatment assignment, then we can at least estimate the average effect of the treatment for the kind of people that choose to remain in the experiment. This is a genuine effect, but we've kind of changed the question. And, of course, most of the time, if there is attrition, we're left worrying that the attrition is both nonrandom and influenced by the treatment. For instance, maybe people leave the study because Body Vibes work so well that they stop worrying about skin health entirely. In that case, were we to compare the remaining members of the treated and untreated groups, we'd be getting a biased estimate of the effect. (This is the sort of thing that can easily happen in a medical study if the researchers aren't careful.) As with many problems, it's much better if you can anticipate and mitigate attrition at the design stage rather than try to account for it after the fact.

If attrition is unavoidable, what should an analyst do? First, you can test whether the experimental treatment influenced the rate of attrition. If it did, then we know we no longer have an apples-to-apples comparison. And relatedly, you can see whether the treated versus untreated units that remained in your sample differ systematically on other covariates that might be related to the outcome.

If you have reason to think the treatment did affect the kinds of respondents that attrited, what can you do? Do we just have to throw out the experiment? Not necessarily—there is one last resort that doesn't require the analyst to make any assumptions about the nature of attrition. You can try to bound the extent of the bias arising from attrition.

To see how this works, imagine an experiment with a binary outcome (1 = healthy skin, 0 = unhealthy skin). Suppose that 50 percent of the subjects in both our treated and untreated groups appear to have healthy skin, suggesting no effect of Body Vibes, but 5 percent of subjects in each group never showed up to have their skin health measured. We don't know whether attrition was affected by the treatment. But we can ask how bad the bias could be if it was.

The best-case scenario for the hypothesis that Body Vibes are good for skin health would be if all of the people in the treated group who didn't show up had good skin and all of the people in the untreated group who didn't show up had bad skin. In that scenario, 52.5 percent of subjects in the treated group would have good skin health compared to 47.5 percent in the untreated group, implying a positive effect of Body Vibes on skin health of 5 percentage points. Alternatively, in the worst-case scenario for this hypothesis, those numbers would be flipped, and there would be a negative effect of 5 percentage points. We can't be sure that attrition doesn't bias our estimates, but we can say that the bias can't possibly be greater than 5 percentage points.

Interference

Interference occurs when the treatment status of one unit affects the outcome of another unit. This can bias the results of an experiment. To see what we mean, consider the following story we heard from our colleague, Chris Blattman, about a pilot study for an experiment he ran in Liberia.

Blattman was interested in understanding what kinds of interventions might help young men at high risk for engaging in crime or violence in post-conflict settings. In particular, he was trying to evaluate the impact of two kinds of interventions: offering young men small cash grants to start an income-generating business and offering them cognitive behavioral therapy.

One thing you might do, if you were interested in whether either of these two approaches works, is to start an organization offering each of them. You could then compare those who received either (or both) of these interventions to those who didn't, to see if those who received them did better in some important way.

Such an approach, however, would fail to compare apples to apples. It could well be that the young men who self-select into receiving grants or therapy are already different from the average young man in the sample. They might be more ambitious, healthier, smarter, or what have you. Thus, it would be a mistake to attribute the entire difference in performance between those who received the grants or therapy and those who didn't to the causal effect of the intervention.

To address these concerns, Blattman designed a randomized experiment in which he randomly assigned the different interventions to different groups of Liberian young men. Everyone would be given a small fee just for participating in the experiment. Then, some participants would get nothing more (the untreated group), while among the remaining participants, some would get a cash grant of about $200, some would get therapy, and some would get both a grant and therapy.

Blattman's plan was to compare levels of crime and homelessness among the young men assigned to different groups. The idea was that if the young men receiving one of the treatments had better outcomes than the untreated group, this would constitute apples-to-apples evidence that the intervention had a positive impact. So far, so good.

The problems started when the young men in the study found out that about half of them would receive $200 while the others would not. They explained that they did not want to play this lottery. They would prefer to each receive $100, eliminating any risk of getting nothing. Of course, giving them each $100 would ruin the experiment. After all, the purpose was to randomly give some more than others and see whether those who received more actually did better. So Blattman's team dispensed the cash grants as per their experimental protocol—randomly giving only half the participants the $200.

But these young men were one step ahead of the researchers. They seemed to have reached an understanding that they would provide one another with a sort of insurance. As a result of this insurance agreement, the winners of the lottery each gave some of their money to the losers, who had received nothing. This kind of interference biased the estimates that came out of the experiment, since now the carefully constructed untreated group had in fact received some of the treatment and the carefully constructed treated group had given up some of the treatment.

Here we see how hard it can be to design a clean experiment. Sometimes your experimental subjects or another outside force will undo your efforts.

Blattman's failed pilot is a clear example of *interference*. When you design an experiment, you randomly assign a treatment of interest across different units of observation (e.g., individual subjects, households, petri dishes). When you do that, you're assuming that those units of observation are independent from one another. However, if the treatment status of one unit actually affects the outcomes of another unit, that's interference, and that can bias the results of your experiment. In this experiment, the interference concern is that the treatment status of the group that got the cash grants affected the outcomes in the untreated group because the treated subjects actually shared some of the treatment with the untreated subjects.

How do careful analysts deal with interference? Sometimes it's interesting enough that the interference itself becomes the object of investigation. Do the taxes in one state influence economic development in a neighboring state? If a campaign mobilizes a group of supporters, will that subsequently mobilize a group of opponents? If a public health program vaccinates children in one school, will this help protect children in another school? Researchers can sometimes design studies with the goal of estimating these kinds of spillover effects. For example, Blattman could have randomly assigned some friend groups to have one person treated with cash and other friend groups to have nobody treated with cash. Then, he could have tested whether the individuals who weren't given money behaved differently when a friend was given money.

In general, careful analysts need to anticipate interference and design their studies in ways that mitigate these possibilities. This is exactly why researchers do things like running pilot studies. In Blattman's case, when he scaled up the experiment after the problematic pilot, he made sure that it remained a secret which subjects had and had not been assigned cash grants, to reduce the risk of interference.

Natural Experiments

For many interesting and important questions, we'd like to learn about causal relationships; however, an experiment might be infeasible, unethical, unrealistic, or prohibitively expensive. But sometimes the world creates something like experimental randomization for us, even without our intervening to actually run an experiment. We already saw one example of this kind of *natural experiment*, in our discussion of the effect of charter schools on academic outcomes in chapter 9. Although no quantitative analyst has been able to conduct their own experiment where they randomly send some kids to charter schools and others to public schools, many charter schools themselves randomize admissions. The schools didn't randomize for scientific reasons but rather because they were required to by law. The law presumably exists because of concerns about fairness and equal opportunity, not causal inference. But regardless of the motivation, these lotteries create randomization "in the wild" that allows us to estimate the

effects of attending charter schools versus regular public schools more credibly than we could by simply comparing the performance of students at the two types of schools and trying to control for all the many potential confounders.

Natural experiments almost always involve some level of noncompliance—for example, not everyone who wins a charter school lottery ultimately attends that charter school, and some people who lose the lottery for one charter school win it for another. Thus, in such settings, we typically either estimate an intent-to-treat effect (i.e., the reduced-form relationship between winning the admissions lottery and academic outcomes) or take an instrumental variables approach to estimate the complier average treatment effect. In this example, the instrument would be winning the lottery, the treatment is attending the charter school, and the outcome is some measure of academic performance (e.g., test scores).

When taking the instrumental variables approach, we need to think seriously about the conditions we described earlier. If there is natural randomization, we can have confidence in exogeneity. That is, we can credibly estimate the effect of winning the admissions lottery on academic performance and on attending the charter school. But we have to think very carefully about the exclusion restriction. That is, are there ways that winning the admissions lottery might affect academic performance other than through its effect on attending a charter school?

It might well be that in the charter schools example, the exclusion restriction is reasonable and that we really can estimate the complier average treatment effect. But let us give you another example where the exclusion restriction is a bit more fraught.

Military Service and Future Earnings

The effect of military service on future earnings is of considerable interest to economists. But, of course, people who serve in the military and do not serve in the military differ in lots of ways that matter for earnings. Hence, a comparison of the earnings of veterans and non-veterans (even controlling for a bunch of stuff) is hopelessly confounded. Such a comparison does not provide a plausibly unbiased estimate of the causal effect.

Fortunately (for social scientists), there is a natural experiment to help. During the Vietnam War, draft-eligible men were randomly assigned draft numbers. People were only actually drafted if their randomly assigned number was sufficiently low. Hence, we have a source of random variation in military service.

Of course, there was not perfect compliance with the draft lottery. For instance, some young men volunteered to serve in the military, despite having a high draft number. (In our earlier terminology, such men are always-takers.) And others, with low draft numbers, left the country or otherwise avoided the draft. (In our earlier terminology, such men are never-takers.) So, if we want to get an estimate of the causal effect of military service on earnings (rather than the reduced-form effect of lottery number on earnings, which seems less interesting), we need to take an instrumental variables approach, which many studies have done. The idea is to use draft number as the instrument, military service as the treatment, and future earnings as the outcome.

In this context, exogeneity is quite plausible. As best we can tell, the government really did assign draft numbers randomly. (Technically, they randomly assigned birthdays, so everyone with the same birthday was in the same boat, but whether one's birthday was selected was random.) So we really can estimate the effect of draft number on military service and on future earnings.

But what about the exclusion restriction? For the exclusion restriction to hold, it needs to be the case that the draft lottery number has no effect on future earnings other than through its effect on military service. How might this be violated?

One possibility concerns how people responded to receiving a low draft number. Such people may have been more likely to engage in various activities that would allow them to avoid the draft. For instance, they may have been more likely to flee the country. Or they may have been more likely to pursue higher education in order to receive a student deferment, which excused them from the draft while they remained in school. Becoming an expatriate or going to college might both directly affect future earnings. As such, these are alternative paths by which the draft number might affect future earnings other than through military service. Because of such violations of the exclusion restriction, it might well be that, even with random assignment of draft numbers, the instrumental variables approach will not allow us to use the draft lottery to credibly estimate the effect of military service on future earnings.

Wrapping Up

There's a reason we call experiments the gold standard for causal inference. By randomly assigning a treatment, we guarantee that the treated and untreated groups have, in expectation, the same potential outcomes, meaning that we can obtain unbiased estimates of a causal relationship.

Even with a randomized experiment, thorny problems can arise. So designing and analyzing experiments requires vigilance and clear thinking. These same thorny problems can rear their heads outside the context of experiments, so we need to continue thinking about them as we move on to other research designs.

Unfortunately for science, the ideal experiment that we'd like to run is often impractical, infeasible, or unethical. What do we do then? The next two chapters discuss special circumstances in which we can still obtain credible estimates of causal relationships even without anything being randomized.

Key Terms

- **Research design:** Approaches to obtaining unbiased estimates of a treatment effect or other estimand.
- **Random assignment:** Deciding which units are assigned to receive treatment in a random fashion (e.g., by flipping a coin or using a random-number generator).
- **Blocked/stratified random assignment:** The process of dividing experimental subjects into different groups (typically groups that you believe have similar potential outcomes) and then randomizing your treatment within each of those groups. This can significantly improve the precision of your estimates. If the probability of treatment varies across blocks or strata, you will have to account for this (e.g., by controlling for block-fixed effects) in order to obtain unbiased estimates.
- **Noncompliance:** When an experimental subject chooses a treatment status other than the one to which it was assigned.
- **Compliers:** Units that take up the treatment status they are assigned.
- **Always-takers:** Units that are always treated, regardless of whether they are assigned to be treated or untreated.

- **Never-takers:** Units that are never treated, regardless of whether they are assigned to be treated or untreated.
- **Defiers:** Units that take up the opposite of the treatment status they are assigned.
- **Intent-to-treat (ITT) or reduced-form effect:** The average effect on the outcome of being assigned to the treated rather than the untreated group. This need not be the average treatment effect because of noncompliance.
- **First-stage effect:** The average effect of being assigned to the treated group on take-up of the treatment. This corresponds to the fraction of compliers.
- **Complier average treatment effect (CATE):** The average treatment effect for the compliers—a special kind of LATE.
- **Instrumental variables (IV):** A set of procedures for estimating the CATE in the presence of noncompliance. The Wald Estimator is a special case of instrumental variables. All IV designs require that we can credibly estimate the effect of the instrument on the treatment and on the outcome (exogeneity), that the instrument affects the treatment (compliers), that the instrument only affects the outcome through its effect on the treatment (exclusion restriction), and that there is not a large number units who take-up treatment if and only if the instrument assigns them to the untreated group (defiers).
- **Exogeneity:** An instrument is exogenous if it is randomly assigned or "as if" randomly assigned such that we can get an unbaised estimate of both the first-stage and reduced-form effects.
- **Exclusion restriction:** An instrument satisfies the exclusion restriction if it affects the outcome only through its effect on the treatment, not through any other channel.
- **Chance imbalance:** The situation where, despite random assignment, the treated and untreated groups differ in important ways because of noise.
- **Statistical power:** The statistical power of a study is technically defined as the probability of rejecting the null hypothesis of no effect if the true effect is of a certain non-zero magnitude. Colloquially, we say that a study has low statistical power if it was unlikely to produce a statistically significant result even if the effect being investigated is large.
- **Attrition:** The situation where experimental subjects drop out of the experiment, such that you do not observe outcomes for those subjects. Attrition is different from noncompliance.
- **Interference:** The situation where the treatment status of one unit affects the outcome of another unit.
- **Natural experiment:** When something was randomized not for research purposes, but careful analysts are nevertheless able to utilize this randomization to answer an interesting causal question.

Exercises

11.1 Suppose a psychology lab attempts to study the phenomenon of behavioral priming. Specifically, they want to know if experimental subjects walk slower when they are exposed to words associated with aging and old age. They recruit subjects to come to their lab and they pay them to complete a word association task. Half the subjects are assigned to an untreated group for which the words have nothing to do with aging, and the other half of the subjects are

assigned to a treated group for which many of the words are related to aging and old age.

After the subjects have completed their task, unbeknownst to the subjects, one of the research assistants times how long it takes them to traverse the fifty-foot hallway that leads to the building's exit. The researchers' plan is to test whether the treatment leads to slower walking times.

Below are some facts about the experiment. For each one, think about what implications that fact has for the experiment. Is this a problem for the researchers? If so, what problem is it? What could they have done in their experimental design or data analysis to address the problem?

(a) The subject pool was a wide cross section of society, so some of the subjects were old, some were young, some were athletic, some were clumsy, some were skinny, some were overweight. The treated group over-represented older and less athletic people, compared to the untreated group.
(b) Some of the subjects didn't pay close attention to the word association activity, gave meaningless answers, and just went through it as quickly as possible.
(c) Some of the subjects took a very long time to walk across the hallway because they stopped to talk to a passerby or to check their phone.
(d) Some of the subjects never crossed the hallway at all because there was another exit through the back of the building.
(e) The research assistants who timed the walking speed of the subjects knew the hypothesis of the researchers and they were the same people who administered the treatments.
(f) Some of the subjects talked to one another about the word association task before they exited the building.

11.2 Download "GOTV_Experiment.csv" and the associated "README.txt," which describes the variables in this data set, at press.princeton.edu/thinking-clearly.

We will be analyzing data from a randomized experiment to estimate the effects of get-out-the-vote (GOTV) interventions on voter turnout.

Several factors complicate the analysis of this particular experiment. First, the probability of being randomly assigned to treatment was different for urban and non-urban areas. Second, some people assigned to treatment did not receive the treatment. And third, we are unable to observe turnout for some of the subjects. See the README file for more details.

(a) Calculate the mean value of turnout for people who did and did not receive the treatment, and interpret the implied effect of get-out-the-vote interventions on turnout. Think about the likely biases that arise from the three complications listed above. If you had to guess, would you say that you are likely over- or under-estimating the average effect with this analysis? Explain your answer.
(b) Using the lessons from chapter 10, try to account for the fact that the probability of treatment varied between urban and non-urban places. How did your estimate change? Why?
(c) Using the lessons from this chapter, let's try to account for noncompliance. First, try to estimate the intent-to-treat effect (reduced form) and

the compliance rate (first stage). Now divide the former by the latter to estimate the complier average treatment effect.

(d) Think about the attrition problem. What are you implicitly assuming if you just drop the subjects for whom we don't observe their turnout? Let's see how our estimates change under different assumptions. Estimate the complier average treatment effect assuming that none of the subjects who attrited would have voted. What would your estimate be under the worst-case scenario for the effectiveness of GOTV? What about the best-case scenario?

Readings and References

For a thorough guide to conducting experiments, particularly field experiments, we recommend

> Alan S. Gerber and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. W. W. Norton.

The study showing that sickly children in Peru were weaned from breastfeeding later is

> Grace S. Marquis, Jean-Pierre Habicht, Claudio Franco, and Robert E. Black. 1997. "Association of Breastfeeding and Stunting in Peruvian Toddlers: An Example of Reverse Causality." *International Journal of Epidemiology* 26(2):349–56.

The randomized experiment on breastfeeding in Belarus is

> Michael S. Kramer, Tong Guo, Robert W. Platt, Stanley Shapiro, Jean-Paul Collet, Beverley Chalmers, Ellen Hodnett, Zinaida Sevkovskaya, Irina Dzikovich, and Irina Vanilovich. 2002. "Breastfeeding and Infant Growth: Biology of Bias?" *Pediatrics* 110(2):343–47.

There are many papers on the Vietnam draft lottery. Two of them (one classic, one recent) are

> Joshua D. Angrist. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80(3):313–36.

> Joshua D. Angrist and Stacey H. Chen. 2011. "Schooling and the Vietnam-era GI Bill: Evidence from the Draft Lottery." *American Economic Journal: Applied Economics* 3(2):96–118.

If the first exercise question made you wonder whether behavioral priming can actually influence someone's walking speed, we recommend the following study. It turns out that the result depends on whether the timing is conducted by a machine or by a human who knows the hypothesis. In other words, it's easy for researchers to trick themselves into thinking they're detecting something when they know what they're supposed to find.

> Stephane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. "Behavioral Priming: It's all in the Mind, but Whose Mind?" *PLoS ONE* 7(1):e29081.