Chapter 16

Causal inference theory and randomized experiments

So far, we have been interpreting regressions predictively: given the values of several inputs, the fitted model allows us to predict y, typically considering the n data points as a simple random sample from a hypothetical infinite "superpopulation" or probability distribution. Then we can make comparisons across different combinations of values for these inputs. This section of the book considers *causal inference*, which concerns what *would happen* to an outcome y as a result of a treatment, intervention, exposure, or other causal variable.

16.1 Basics of causal inference

For most statisticians, ourselves included, causal effects are conceptualized as a comparison between different potential outcomes of what might have occurred under difference scenarios. This comparison could be between a factual state (what did happen) and one or more counterfactual states (representing what might have happened), or it could be a comparison among various counterfactuals.

Running example

In this chapter we shall use an example of a hypothetical small experiment on the effect of a dietary intervention. By using artificial data we will be able to demonstrate many of the key ideas of causal inference. Unobserved potential outcomes are a key aspect of causal inference, and with fake data we can look at all these potential outcomes and better understand how they are combined to define different causal summaries.

Increasingly over the past two decades, consumption of omega-3 fatty acids has been touted as an effective strategy for addressing a variety of medical conditions ranging from coronary heart disease to rheumatoid arthritis to anxiety and depression. Suppose you had done some reading on the subject and were concerned about your own blood pressure levels but were confused by the conflicting evidence regarding the effect of omega-3 fish oil supplements on blood pressure. So you decided to investigate the link yourself. You found 8 friends who agreed to be part of an informal study on the relationship between fish oil supplements and systolic blood pressure. Four of the friends were placed in the "fish oil supplement" treatment group. Members of this group agreed to consume 3 grams of fish oil supplements per day for one year while otherwise maintaining their current diets. The other four friends agreed to simply maintain their current diets free from fish oil supplements for the same year. At the end of the study period blood pressure was measured for each of the eight participants. To simplify the discussion (and recognizing that classifications of hypertension vary) we will henceforth consider only systolic blood pressure, and consider pressure of 160 and above to represent "high blood pressure." We also assume that study participants actually receive the treatment they are assigned. Before we examine the study

design and data more closely, however, let us take a step back and be explicit about what we are trying to estimate.

Potential outcomes, counterfactuals, and causal effects

To formalize this notion we start by creating notation for the treatment variable that takes one of two values. In our running example, $z = 0$ denotes "no fish oil supplements ingested" and $z = 1$ denotes "3 grams per day of fish oil supplements ingested." To think about what it means for fish oil supplements (relative to no supplements) to *cause* lower blood pressure, we need to consider two possible outcomes: the blood pressure that would result if the person had no supplement, y_i^0, and the blood pressure that would result if the person had received the prescribed supplement, y_i^1. These possible outcomes are commonly referred to in the causal inference literature as *potential outcomes*. For the people in the experiment, the "factual state" is represented by the potential outcome that corresponds to the treatment actually received (y_i^1 for those who received the treatment and y_i^0 for those who did not). The "counterfactual state" is represented by the potential outcome that corresponds to the treatment not received (y_i^0 for those who received the treatment and y_i^1 for those who did not). The counterfactual state is thus not observed and only could be if we went back in time and had the study participant choose the opposite treatment. For the people not in the experiment, both states are counterfactual and we are interested in what might happen under either scenario.

The observed outcome for the people in the experiment is simply the potential outcome that was revealed due to treatment assignment and thus can be considered to be a function of both potential outcomes, $y_i = y_i^0(1-z_i)+y_i^1 z_i$. The causal effect of supplements versus no supplements for person i, can be expressed by a mathematical statement such as $\tau_i = y_i^1 - y_i^0$. Causal effects can also be expressed as nonlinear functions of the potential outcomes. In this book we focus on linear functions because they are conceptually and mathematically simpler and because linear models work well in many settings, especially after appropriate transformations.

Why go through all the effort of defining potential outcomes and counterfactual states? Consider the clarity that is gained with this approach. Suppose the j^{th} individual, Audrey, was observed pre-treatment to have a blood pressure of 140. She received supplements for a year at which time her systolic blood was measured at 135; thus $y_j^1 = 135$. (We assume throughout that measurement error is small enough compared to natural variation that it can be ignored.) Does the pre/post decline from 140 to 135 provide evidence that fish oil supplements *caused* a reduction in Audrey's blood pressure? The only way we could make this claim would be if we knew that *if* Audrey had not received the supplements, then her blood pressure *would have been* higher, for instance if we knew that $y_j^0 = 140$ for her. If, instead, Audrey would have had the same blood pressure either way—that is, if $y_j^0 = 135$ for Audrey as well—then the causal effect of supplements on her blood pressure would be 0.

The fundamental problem of causal inference

The problem inherent in determining the causal effect for any given individual, however, is that we can never observe both potential outcomes y_j^0 *and* y_j^1. In the language of this example we cannot observe the blood pressure that would have resulted *both* if Audrey had taken the supplements *and* if she had not, thus the causal effect is impossible to directly measure. This is commonly referred to as the *fundamental problem of causal inference*.

To see this in a broader context consider the eight person study you implemented. Figure 16.1 displays the causal inference data that the researcher could observe. Even though we observe outcomes that manifest under each of the two treatment regimes, we cannot observe any given person's outcome under both regimes. Not only can we not determine

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	0	140	?	140
Anna	1	40	0	140	?	140
Bob	0	50	0	150	?	150
Bill	0	50	0	150	?	150
Caitlin	1	60	1	?	155	155
Cara	1	60	1	?	155	155
Dave	0	70	1	?	160	160
Doug	0	70	1	?	160	160

Figure 16.1 *Hypothetical causal inference data for the effect of fish oil supplements on systolic blood pressure. For each person, we as researchers can only observe the potential outcome corresponding to the treatment actually received. Therefore we cannot directly observe either the individual-level or group-level causal effects. In this example, a simple difference of average outcomes across the two groups,* $157.5 - 145$*, would lead to the estimate that supplements produce a 12.5 mmHg increase in systolic blood pressure. This is a poor estimate because the treatment and control groups here are highly imbalanced.*

any individual causal effect we cannot even determine an average causal effect across the eight study participants (without further assumptions)!

Close substitutes

But what if we had Audrey's pre-study blood pressure? Could we use that as a substitute for her y_j^0? This is the basic idea behind a "pre-post" or before-after study design. The problem is that Audrey's blood pressure before the study starts, y_j^{before}, is not necessarily an accurate reflection of what her blood pressure would be without the supplements one year later. For all we know Audrey might have had other life changes during that study year: she might have lost her job or started a new one, she might have gotten married or divorced, she might have taken up running or discovered a love of baby back ribs. Any of these life changes could have affected her blood pressure irrespective of whether she received the omega-3 supplements or not, making y_j^{before} a poor substitute for y_j^0. In Section 17.4 we discuss before-after studies in the context of a particular example.

Things would get still more complicated if the study were implemented such that Audrey received the supplements in the first year but not the second. If the supplements turned out to have a long-lasting effect, then the results from the second year would be muddied by the treatment effect from the first year. We might attempt to avoid this contamination by allowing for a wash-out period between study segments (in our example to attempt to make sure the body is no longer being affected by the supplements), but this design still suffers from the defect of implicitly attempting to substitute measurements from one time period as estimates of potential outcomes from another.

There are ways of adding rigor to these types of pre-post designs. For instance one could include many participants, randomize the order of receipt of treatment, and take multiple measurements on each person. Such studies fall under the general classification of *crossover trials*.

It is not unusual to see studies that attempt to make causal inferences by substituting values in this way. It is important to keep in mind the strong assumptions often implicit in such strategies which we can now formalize by explicitly requiring that y_j^0 and y_j^1 correspond to the same point in time. In the absence of this correspondence other confounding factors might influence the potential outcomes, creating differences between them attributable to factors other than just the treatment.

More pristine examples can generally be found in the natural and physical sciences. For

instance, imagine dividing a piece of plastic into two parts and then exposing each piece to a corrosive chemical. In this case, the hidden assumption is that pieces are identical in how they would respond with and without treatment, that is, $y_1^0 = y_2^0$ and $y_1^1 = y_2^1$. In experiments with humans or animals, one must typically either assume that the effect of the initial treatment goes away before the next treatment is applied, or else one must fit some sort of model to the carryover.

More than two treatment levels, continuous treatments, and multiple treatment factors

Going beyond a simple treatment-and-control setting, multiple treatment effects can be defined relative to a baseline level. With random assignment, this simply follows general principles of regression modeling.

If treatment levels are numeric, the treatment level can be considered as a continuous input variable. To conceptualize randomization with a continuous treatment, think of spinning a spinner that can land on any of the potential levels of the treatment assignment. As with regression inputs in general, it can make sense to fit more complicated models as suggested by theory or supported by data. A linear model—which estimates the average effect on y for each additional unit of z—is a natural starting point for effects that are believed to be monotonically increasing or decreasing functions of the treatment level.

With several discrete treatments that are unordered, as in a comparison of three different sorts of psychotherapy, we can move to multilevel modeling, with the group index indicating the treatment assigned to each unit, and a second-level model on the group coefficients, or treatment effects. We discuss such models in the next volume.

Additionally, multiple treatments can be administered in combination. For instance, depressed individuals could be randomly assigned to receive nothing, drugs, counseling sessions, or both drugs and counseling sessions. These combinations could be modeled as two treatments and their interaction or as four distinct treatments.

16.2 Average causal effects

The counterfactual or potential outcomes framework adds clarity to the meaning of causal effects, but it also highlights the challenge inherent in estimating them. In experiments in which only one treatment is applied to each individual, it will not be possible to estimate individual-level causal effects. So how do we proceed? Can we gain any traction by collecting data on many individuals, some who took the supplements and some who did not?

Consider again our hypothetical eight-person study. Suppose Figure 16.2 were a display of the data that could be seen if we were somehow able to apply both the treatment and the control to each person. Under this state of omniscience we would get to see both y_i^1 and y_i^0 for each study participant, even though the researcher in any real study could see at most one of these for each participant. Given all these potential outcomes, we can directly compute the treatment effect for each person:

$$\text{individual treatment effect: } \tau_i = y_i^1 - y_i^0.$$

For the example shown in Figure 16.2, the treatment effect for each of the men is a reduction in systolic blood pressure of 10 mmHg; the effect for each of the women is a reduction of 5 mmHg.

The value of framing this as a causal inference question, rather than simply comparing observed outcomes, is that when there are differences between treatment and control groups (as there will be in practice), some adjustment will be needed to estimate the causal effect—and for this it helps to understand what underlying summary is being estimated.

The causal effect, $y_i^1 - y_i^0$, can in general vary by person, hence any definition of average

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	0	**140**	135	140
Anna	1	40	0	**140**	135	140
Bob	0	50	0	**150**	140	150
Bill	0	50	0	**150**	140	150
Caitlin	1	60	1	160	**155**	155
Cara	1	60	1	160	**155**	155
Dave	0	70	1	170	**160**	160
Doug	0	70	1	170	**160**	160

Figure 16.2 *Hypothetical causal inference scenario for the effect of fish oil supplements on systolic blood pressure. Based on these numbers, there is a treatment effect of* -10 *for men and a treatment effect of* -5 *for women. This table includes all potential outcomes including those that would not be observed in reality. For each row, the observation in bold is the one actually seen, and the regular font observations could never be known. A naive comparison of outcomes across the two groups* $(157.5 - 145)$ *would lead to the erroneous conclusion that supplements produce a 12.5 mmHg* increase *in systolic blood pressure.*

causal effect will depend on what group of people is being averaged over. This is similar to the choice of population distribution in poststratification, as discussed in Section 15.1, but here we will demonstrate in our simple example with just eight individuals.

Sample average treatment effect (SATE)

For the data shown in Figure 16.2, the average treatment effect across all eight units in this example is a 7.5 mm Hg reduction in systolic blood pressure. This estimand, called the *sample average treatment effect*, or SATE, can be calculated by averaging all the y^1's in the sample, $\frac{1}{n}\sum_{i=1}^n y_i^1$, and subtracting the average of the y^0's in the sample, $\frac{1}{n}\sum_{i=1}^n y_i^0$. Equivalently, and perhaps more intuitively, we can simply average the individual-level causal effects,

$$\tau_{\text{SATE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0).$$

Conditional average treatment effect (CATE)

We can also calculate the average treatment effect for well-defined subsets of the sample such as men (for which average treatment effect in this example is -10, as can be seen by averaging the relevant numbers in Figure 16.2), women (average treatment effect of -5), or, for instance, 50-year-olds (average treatment effect of -10). These estimands are sometimes referred to as *conditional average treatment effects* (CATEs) and can also take more complicated forms such as expectations (average predictions) from linear regression models.

Population average treatment effect (PATE)

Researchers often want inferences about some *population* of interest rather than simply the study sample. Realistically, the people (or, more generally, experimental units) in a social or medical experiment are typically not a representative sample of any known or well-understood population of inference. In fact, often the more well-tailored a sample is for making causal inferences for a given sample, the less likely the sample is to be a random sample from or representative of a known population. For instance, it is quite rare for a

randomized experiment involving human subjects to comprise any type of random sample from a known or well-defined population.

Nonetheless, the *population average treatment effect* (PATE) is still often considered to be an inferential goal of causal inference. For a population of size N we could then define the PATE as

$$\tau_{\text{PATE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i^1 - y_i^0).$$

Therefore to *know* the PATE our omniscience would need to extend to seeing both potential outcomes for the full population of interest. In essence, in this inferential problem, there are then two types of missing data: (1) the missing counterfactual values within our sample, and (2) both potential outcomes for those observations not in our sample.

If our study sample is a random sample of the population of interest, then any unbiased estimate of SATE will also be an unbiased estimate of PATE. More generally, when we fit regression models, the estimate of PATE will depend on the assumed distribution of the regression predictors in the population of interest. For example, if the estimated treatment effect is larger among women than men, then the PATE will depend on the proportion of men and women in the population.

In the absence of a random sample, *estimating* the PATE requires a model of the treatment effect given pre-treatment predictors that will allow us to extrapolate from experimental units to the general population, in the same way that poststratification allows us to generalize from imperfect samples as discussed in Section 15.1. We discuss such methods in the section on generalization.

Problems with self-selection into treatment groups

Let's return to our inferential goal of trying to estimate an average treatment effect. The starting point is the comparison of treatment and control groups, and we can run into trouble if these two groups are not sufficiently similar or *balanced*. In Figure 16.1, we see that the people who received treatment were on average older than the controls. This could be just by chance or perhaps because those who agreed to take the supplements were more concerned about their blood pressure and the study offered them a chance to try out the supplements for free, while those who agreed to be in the no-supplements group did not care one way or the other whether the supplements might benefit their health.

What is the implication of this type of sorting or self-selection into treatment groups? In our hypothetical example we know all the counterfactual outcomes, and we have already seen from Figure 16.2 that the true sample average treatment effect is -7.5. However, what happens if we perform the most straightforward analysis and simply compare the average outcomes for those who chose to be in the treatment group to the average outcomes for those who chose to be in the control group? The difference in means between these two groups in Figure 16.1 yields a naive estimated treatment effect estimate of -12.5!. What went wrong?

Applied researchers will intuitively know the answer to this question. The groups whose outcomes we were comparing differed in their pre-treatment characteristics. Although we had the same ratio of men to women across the two groups, those in the control group were quite a bit younger on average than those in the treatment group. This difference matters because age is also predictive of the outcome.

Perhaps more interestingly, even if the observed characteristics had been exactly the same across the two groups, if the potential outcomes remained as they are in Figure 16.2 clearly the simple comparison would have given the wrong answer. In fact, *the potential outcomes encapsulate all of the necessary information regarding what we need to be similar across the two groups*. Since we can't directly adjust for these though, we use our pre-treatment variables as a proxy.

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	0	**140**	135	140
Anna	1	40	1	140	**135**	135
Bob	0	50	0	**150**	140	150
Bill	0	50	1	150	**140**	140
Caitlin	1	60	0	**160**	155	160
Cara	1	60	1	160	**155**	155
Dave	0	70	0	**170**	160	170
Doug	0	70	1	170	**160**	160

Figure 16.3 *Hypothetical causal inference data for the effect of a fish oil supplements on systolic blood pressure from an idealized randomized experiment that happens to achieve perfect balance. There is a treatment effect of* -10 *for men and a treatment effect of* -5 *for women. For each row, the observation in bold is the one actually seen. In practice, the regular font-weight observations would not be known. In this case the randomization got lucky and the simple difference in means,* $155 - 147.5$*, happens to recover the true sample average treatment effect (SATE) of* -7.5*. Compare, however, to Figure 16.4, which shows how the same design can give a much different result in this small sample.*

Using design and analysis to addressing imbalance between treatment and control groups

In practice we can never ensure that treatment and control groups are balanced on all relevant pre-treatment characteristics. However there are statistical approaches that may bring us closer. At the design stage, we can use *randomization* to ensure that treatment and control groups are balanced in expectation, and we can use *blocking* to reduce the variation in any imbalance. At the analysis stage, we can *adjust* for pre-treatment variables to correct for differences between the two groups to reduce bias in our estimate of the SATE. We can further adjust for differences between sample and population if our goal is to estimate PATE.

16.3 Randomized experiments

Randomly assigning units to treatment and control groups ensures that there are no differences in expectation in the distribution of potential outcomes between groups receiving different treatments (or treatment and control). We can randomly assign units to treatments (or treatments to units) by flipping a coin, drawing names from a hat, or, more generally, assigning random numbers to units. In this day and age we usually let a computer software package create random assignments. For instance in our example the following command in R could be used to create a randomly ordered vector, named z, of four ones and four zeroes:

```
z <- sample(c(0,0,0,0,1,1,1,1), 8, replace=FALSE)
```
R code

Alternately, the participants could be assigned numbers from 1 to 8 and then four of these numbers (say 1 to 4) could be randomly chosen to indicate which subjects should receive the treatment, as in:

```
z_ind <- sample(1:8, 4, replace=FALSE)
z <- as.numeric(z_ind < 5)
```
R code

Completely randomized experiments

In a completely randomized experiment, the probability of being assigned to the treatment is the same for each unit in the sample. Let us go back in time and randomly assign your eight friends to treatment conditions.

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	1	140	**135**	135
Anna	1	40	1	140	**135**	135
Bob	0	50	1	150	**140**	140
Bill	0	50	0	**150**	140	150
Caitlin	1	60	0	**160**	155	160
Cara	1	60	0	**160**	155	160
Dave	0	70	0	**170**	160	170
Doug	0	70	1	170	**160**	160

Figure 16.4 *Hypothetical causal inference data for the effect of fish oil supplements on systolic blood pressure from a randomized experiment that happened to have a less perfectly balanced assignment, compared to that shown in Figure 16.3. The difference in means is again an unbiased estimate of the treatment effect under this completely randomized design, but it is subject to sampling variation and for this particular realization the estimate,* $142.5 - 160 = -17.5$*, happens to be far from the underlying and unobservable sample average treatment effect (SATE) of* -7.5*.*

An idealized outcome from a completely randomized experiment. Figure 16.3 displays the same units as in Figures 16.1–16.2, but now the treatment assignment and observed outcome reflect an idealized random assignment. As with Figure 16.2 we are omniscient, and thus both potential outcomes are revealed for each unit. We refer to the random assignment as idealized because it happens to have led to perfectly balanced distributions of potential outcomes.

In this case, given the perfect balance across groups, it makes sense that a simple difference in means exactly recovers the true treatment effect of -7.5. Equivalently, the average causal effect of the treatment corresponds to the coefficient τ in the regression, $y_i = \alpha + \tau z_i + \text{error}_i$, thus we could estimate the treatment effect using a regression of the outcome on the treatment assignment, as we discuss in the next chapter.

A less idealized randomized result. What happens if our randomization yields treatment assignments that are not so perfectly balanced? Consider for example the treatment assignment in Figure 16.4. In this case the random assignment led to imbalance between treatment and control groups (see Exercise 16.** for insight into the probability of seeing this level of imbalance in a randomized experiment). In this particular case, the younger participants were disproportionately more likely to have received treatment. More to the point, the y_i^0's are 140, 140, 150, and 170 in the control group and 150, 160, 160, and 170 in the treatment group. This implies that even if the treatment had no effect at all, we'd see a big difference between treatment and control groups in the measured outcome.

For this particular treatment assignment, the simple difference between average treatment and control measurements comes to, -17.5, which is quite a bit off from the true sample average treatment effect of -7.5. Randomization ensures balance on average but not in any given sample, and imbalance can be large when sample size is small.

16.4 Sampling distributions, randomization distributions, and unbiased estimation

An unbiased estimate leads us to the right answer *on average.* What does that mean? In classical statistical inference we think about using samples to estimate population quantities. Properties of a statistical procedure are reflected in the distribution of the estimate over repeated samples from that population. That is, we can envision taking an infinite number of samples from the population and for each sample calculating the estimate. The distribution of these estimates is the sampling distribution.

The estimate is *unbiased* if the mean of this sampling distribution is equal to the estimand. The estimate is *efficient* if the sampling distribution has small variance. How can we conceptualize a sampling distribution when we are estimating the sample average treatment effect from the sample itself?

As an alternative way to conceptualize the inherent variability in our estimate, imagine a different way to think about the uncertainty. First of all, we can simplify matters by considering all covariates and potential outcomes to be fixed (this is a representation common both to the survey sampling world and the randomization-based inference framework). Then imagine randomly allocating observations to treatment groups again and again. Importantly, each new allocation will imply a different set of observed outcomes (since observed outcomes are a function of both potential outcomes and treatment assignment). Suppose that with each re-randomization the difference in mean outcomes between the treatment and control groups is calculated. For the k^{th} sample we could write this as $d^k = \frac{1}{n_1}\sum_{i;z_i^k=1} y_i^k - \frac{1}{n_0}\sum_{i;z_i^k=0} y_i^k$. The set of these estimates represents the *randomization distribution* for this estimator. If the estimator is unbiased then the average of all of these estimates (the mean of the randomization distribution) would equal the true sample average treatment effect; that is, $\mathrm{E}(d_k) = \tau_{\mathrm{SATE}}$.

When considering the population average treatment effect, we can conceptualize drawing a random sample of size n_1 from a population to receive the treatment, drawing a random sample of size n_0 from the same population to receive the control, and taking the difference between the average response in each as an estimate of the population average treatment effect. If this were done over and over, the estimates that were yielded would form a sampling distribution for this estimate. For the estimate to be unbiased we would need the mean of this sampling distribution to be centered on the true τ_{PATE}. A stricter criterion that would also lead to an unbiased treatment effect estimate would require the sampling distribution of the mean control y's to be centered on the population average of y^0 and the sampling distribution of the mean treated y's to be centered on the population average of y_1.

While unbiasedness can be an important quality for an estimate, it does not guarantee that the estimate from any particular random assignment will be close to the true value of the estimand, particularly when the sample is small. The sampling or randomization distribution may be wide, reflecting unlucky, but still possible, randomizations such as the one represented in (16.4). One way we can increase the chances that our estimate is closer to the true value of the estimand is through design choices that reduce the potential for imbalance and thus decrease the variance of the treatment effect estimates. Exercise 16.4 provides the opportunity to simulate randomization distributions.

16.5 More complicated randomized designs

Consider once again the data in Figure 16.3 that resulted from the idealized assignment with perfect balance. Let us look more closely at what made that configuration so successful at recovering the true sample average treatment effect. There are in essence four different types of participants in the study. One way to characterize these four types would be: women aged 40, men aged 50, women aged 60, and men aged 70. A more definitive characterization would reference the four unique pairs of potential outcomes, y_i^0 and y_i^1: 140 and 135, 150 and 140, 160 and 155, and 170 and 160. Let's focus on the observable characteristics however since a researcher could observe those.

We will now refer to these groups as *blocks*. As described above, given that there are equal numbers of control and treated units within each block, each treated unit in the block gets to act as the perfect comparison for the control unit and vice versa. In this idealized example there is not even sampling variability: the potential outcomes are exactly equal, not simply drawn from the same distribution. Therefore the distribution of y^0's in the treatment

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	0	**140**	135	140
Abigail	1	40	0	**140**	135	140
Arielle	1	40	0	**140**	135	140
Anna	1	40	1	140	**135**	135
Bob	0	50	0	**150**	140	150
Bill	0	50	0	**150**	140	150
Burt	0	50	0	**150**	140	150
Brad	0	50	1	150	**140**	140
Caitlin	1	60	0	**160**	155	160
Cara	1	60	1	160	**155**	155
Cassie	1	60	1	160	**155**	155
Cindy	1	60	1	160	**155**	155
Dave	0	70	0	**170**	160	170
Doug	0	70	1	170	**160**	160
Dylan	0	70	1	170	**160**	160
Derik	0	70	1	170	**160**	160

Figure 16.5 *Hypothetical causal inference data for the effect of fish oil supplements on systolic blood pressure from a* randomized block experiment. *Compare to Figures 16.4 and 16.4, which show the same pre-treatment predictors and the same potential outcomes but with different patterns of treatment assignments. With the randomized block design shown here, a simple difference in means,* $152.5 - 150 = 2.5$*, is no longer an unbiased estimate of the true SATE of* -7.5*, but an appropriately adjusted estimate will be unbiased.*

group is exactly the same as the distribution of y^0's in the control group. The same is true for the y^1's.

Randomized block experiments

What would have happened if the ratio of treated to control subjects were not equal across blocks? Consider the data displayed in Figure 16.5 arising from a design involving eight extra recruits from among your friends. The number of treated units is no longer equal to the number of control units within each block. In fact, the ratio of treated to control units now varies across blocks. In each of the top two blocks in the table, there are three controls and one treated. The bottom two blocks contain one control and three treated units. Because of this, the overall distribution of y^0's in the treatment group is quite different from the overall distribution of y^0's for the controls. This level of imbalance in potential outcomes between the treatment and control groups might occur in such a small sample with a completely randomized design although it would be unlikely (see Exercise 16.4 at the end of this chapter to quantify this). In this scenario the simple difference in means is not close to the true SATE either for this sample or across repeated samples.

How did the data from this version of the hypothetical study arise? That is, what was the study design? Suppose the older participants had objected to a completely randomized design because they thought that they should be given preference regarding access to the supplements. Wanting to address these concerns but still feeling committed to randomization, you could have decided to give the eight oldest participants preference by allowing them a higher probability of receiving the supplements. Within each age group you might also have considered it important to ensure there were at least one male and one female in each treatment condition. These considerations led to a design in which randomization was performed within each group defined by age and sex (or first letter of the first name as it so happens). In the first two blocks (Audrey through Brad) that contain the younger

participants, the probability of receiving the supplements is 0.25 under this design, whereas for the last two blocks, with the older participants, this probability is 0.75.

Now that we know that observations were randomized within blocks, we can estimate average treatment effects, *taking into account the design in the analysis*. In practice this means that we calculate treatment effects conditional on the blocks within which the randomization was performed, and then average over the blocks to represent the sample, or the general population. We can do this in any one of several ways.

The simplest approach conceptually starts with calculating a separate treatment effect for each block, $\hat{\tau}_j$. Since this design implies that a separate (tiny) randomized experiment was implemented within each of these blocks of units, each of these is a valid (if potentially imprecise) treatment effect estimate for that block. To get an estimate of the sample or population average treatment effect (SATE or PATE) we could just average these estimates with weights proportional to the number of units in each, specifically $\sum_j n_j \hat{\tau}_j / \sum_j n_j$ for the SATE or $\sum_j N_j \hat{\tau}_j / \sum_j N_j$ for the PATE.

Another approach is to fit a linear regression of blood pressure on the treatment variable and indicators for the three (of the four) blocks (denoted by b^k),

$$y_i = \alpha + \tau_{\text{RB}} z_i + \gamma_1 b_i^1 + \gamma_2 b_i^2 + \gamma_3 b_i^3 + \text{error}_i.$$

where the coefficient of interest is τ_{RB}. We might be able to sharpen this estimate by adding sex as a predictor. In theory we could also recover block-specific treatment effect estimates if we include block by treatment interactions in the model. We demonstrate R code for these approaches in Sections 17.3–17.5.

Motivations for randomization in blocks. In this example the randomized block experimental design was chosen in part to increase people's willingness to participate in the study. Given that the older participants might have had more to gain from taking the supplements, this could be thought of as a design motivated by both logistical and ethical concerns. Another logistical constraint that often motivates use of randomized block experiments is location. When an experiment is being implemented across several different geographical sites, it can be simpler to implement the randomization separately within site.

The motivation for a randomized block design may also be purely statistical, since when implemented successfully it can reduce the uncertainty in the estimate of the treatment effect. This can be particularly important when measurements are noisy or sample size is small.

Defining blocks in practice. How do we choose blocks when our goal is to increase statistical efficiency? The goal when creating blocks is to minimize the variation of each type of potential outcome, y^0 and y^1, within the block. In our contrived example we have access to data on both potential outcomes for each person and thus in theory could identify blocks within which there is no variation in the values of either potential outcome across participants. In practice researchers only have access to observed pre-treatment variables when making decisions regarding how to define blocks. The randomized block experiment in this section is an example with blocks defined by age. So, to the extent possible, the predictors used to define the blocks should be those that are believed to be predictive of the outcome based on either theory or on results from previous studies. The more predictive the blocking variable, the bigger the precision gains at the end of the day.

Matched pairs experiments

Suppose that there is an important characteristic of your friends who participated in this study that you neglected to reveal at the outset. The 16 friends are actually a collection of 8 identical twins! We can identify these pairs because in each case the siblings have names that start with the same letter of the alphabet. This feature of the sample presents a new

design possibility: a matched pairs randomized experiment. This design is a special case of a randomized block design in which each matched pair represents a 2-unit block. The randomization that occurs simply assigns one member of each pair to receive the treatment and the other to receive the control.

If the members of the pair are closely matched to each other this design can yield substantial efficiency gains. In fact the data presented in Figure 16.4 would be much more likely to arise from a matched pairs randomized design than from a completely randomized design, and we have discussed the idealized nature of this design that arises from the fact that it can be characterized as sets of identically matched pairs.

In practice it can be difficult to find pairs that are closely matched even on their observable characteristics, let alone their potential outcomes, which are unobserved at the time of treatment assignment. If the members of the pair do not have similar potential outcomes, the paired design will not reduce variance in the estimated treatment effect.

Often the most effective paired or blocked designs arise in situations in which pairs or groupings arise naturally in the data, such as multiple children in a family or multiple students in a school. In general an effective strategy can be to match based on a wide variety of characteristics using a dimension-reduction strategy such as Mahalanobis distance or propensity score matching, as we discuss in Chapter 18.

Group or cluster randomized experiments

Sometimes it is difficult or problematic to randomize treatment assignments to individual units and so researchers instead randomize groups or clusters of observations to receive the treatment. This design choice might be motivated by the nature of how a treatment is implemented. For instance, group therapy strategies typically need to be implemented at the group level—that is, with every individual in the group receiving the same treatment. Similarly school-wide reform interventions by definition require treatment assignment at the school level.

A decision to assign treatments at the group level can be driven by cost or logistical concerns. It might be more cost-effective to provide free flu shots to a random subset of health clinics, for example, than to have professionals go to every clinic and then randomly assign individuals to receive shots. Assignment at the clinic level would also avoid potentially creating ill-will among potential study participants being deprived of a service that others in the same location are able to receive. Cluster randomized experiments are also used as a strategy to avoid problems caused by spillover or contagion effects (also referred to as SUTVA violations as discussed later).

When treatments have been assigned at the group level, the simplest approach to analysis is to conceptualize the groups as the level of analysis and use aggregated measures of the response variables as the outcome in the analysis. The example discussed later in this chapter does exactly this. An alternative is to use multilevel regression to model the outcomes at the individual level while including treatment assignment as a group-level predictor; we discuss further in the second volume.

16.6 Properties, assumptions, and limitations of randomized experiments

By examining the consequences of our design choices, we have revealed some desirable properties of the randomized experiment. When the design yields no differences, on average, between groups, a simple difference in means will yield an unbiased estimate of the sample average treatment effect, and a regression on treatment indicator, also controlling for pre-treatment predictors, will also be unbiased but can do even better by reducing variance. This section provides a more detailed discussion of the properties implied by randomization

for completely randomized experiments, randomized block experiments, and matched pairs designs.

Ignorability

We first discuss how the ignorability assumption differs across different randomized study designs.

Completely randomized experiments. In a completely randomized design, the treatment assignment is a random variable that is independent of the potential outcomes, a statement that can be written formally as,

$$z \perp y^0, y^1. \tag{16.1}$$

The implication of this independence is that, under repeated randomizations, there will be no differences, on average, in the potential outcomes, comparing treatment and control groups. This property is commonly referred to as *ignorability* in the statistics literature (though it is a special case of a more general assumption also referred to as ignorability as we shall see shortly). Ignorability does not imply that the groups are perfectly balanced. Rather it implies that there is no imbalance *on average* across repeated randomizations. If we were to redo the randomized assignment we'd be equally likely to see imbalance in one direction as in another. Said another way, ignorability implies that the value of someone's potential outcomes does not provide any information about his or her treatment group assignment. For instance, someone who would have low blood pressure under either regime would not be any more likely to end up in the treatment group.

More formally, the independence between the treatment indicator z and the potential outcomes that exists under this assumption implies the unbiasedness property discussed in Section 16.4. To see why, consider the following equivalencies:

$$\begin{aligned} \mathrm{E}(y|z=1) &= \mathrm{E}(y^1|z=1) = \mathrm{E}(y^1) \\ \mathrm{E}(y|z=0) &= \mathrm{E}(y^0|z=0) = \mathrm{E}(y^0). \end{aligned}$$

The first line says that the average of the outcomes for those assigned to treatment is equal to the average of the treatment potential outcomes for those assigned to treatment which is equal to the average of all the treatment potential outcomes. The first equivalency in this statement is achieved by definition; observed outcomes for those in the treatment group are y_i^1's. The second equivalency holds due to ignorability. Since y_1 and z are independent, conditioning on z makes no difference in expectation. Parallel arguments can be made for the second statement in this pair. Put together we can see that the observed difference in means across treatment and control groups will be an unbiased estimate of the true treatment effect.

In this discussion we do not specify if the expectation is over the sample or the population. To consider the former, think about the randomization distribution discussed above. These properties imply that it will be centered on the true SATE. If the data are also a random sample from the population then they additionally imply that the sampling distribution will be centered on the true PATE.

Another important property of randomized experiments is that they create independence between treatment assignment and *all* variables x that occur before treatment assignment, so that $x \perp z$. In general, the relevant cutoff time is when the treatment is *assigned*, not when it is *implemented*, because people can adjust their behavior based on the anticipation of a treatment to which they have been assigned. To put it another way, if the act of assignment is potentially part of the treatment, then the study should be analyzed as such.

Pre-treatment variables include those that do not change value over time or variables such as participant age, which change over time but cannot be affected by a treatment. We

can think of the potential outcomes as a subset of these pre-treatment variables because we can conceptualize them as existing before the treatment is even implemented, as shown in Figure 17.1 on page 332. The treatment assignment identifies which of the potential outcomes we will get to observe.

This property is what motivates researchers who implement randomized experiments to check whether the randomization worked well by comparing means or distributions of observed pre-treatment covariates across randomized treatment groups. It is also important because it justifies use of subgroup difference in means (and regression analogs of such estimates) as unbiased estimates of conditional treatment effects. It also justifies inclusion of pre-treatment variables as predictors in models that estimate treatment effects using data from randomized experiments (more below).

Randomized block experiments. When we randomize within blocks, the independence between potential outcomes and the treatment variable holds only conditional on block membership, w:

$$z \perp y^0, y^1 \,|\, w. \tag{16.2}$$

This more general version of the randomization property is also known as ignorability (sometimes called conditional ignorability, but the word "conditional" here is redundant), as will any such statement of conditional independence between the potential outcomes and the treatment assignment as long as the data being conditioned on (here w) are fully observed. Intuitively we can think of a randomized block experiment as a collection of experiments each of which are performed within the members of a specific block. This means that all those in the same block (those who look the same with regard to their w variable) have the same probability of being assigned to the treatment group.

The implication of this property is that it is necessary to condition on the blocks when estimating treatment effects as described in the previous section. The caveat to this rule is that if the probability of receiving the treatment is the same across all the blocks we can ignore this conditioning and still get an unbiased treatment effect estimate. However to achieve the efficiency gains of this design the conditioning is still necessary.

Matched pairs experiments. Recall that the paired design is the special case of randomized blocks in which each block contains exactly two units. Therefore we can formalize the randomization property for matched pairs in the same way as in (16.2), where w now indexes the pairs. Given that there are only two units in each block of the paired design and exactly one must receive the treatment, the probability of receiving the treatment is, by definition, the same across blocks. Therefore a simple difference in means will constitute an unbiased treatment effect estimate in this setting. However, this strategy will miss out on the potential efficiency gains of the design.

The most straightforward approach to treatment effect estimation that accounts for the matched pairs design is to calculate pair-specific differences in outcomes, $d_j = y_j^T - y_j^C$ (where y_j^T is the outcome of the treated unit in pair j and y_j^C is the outcome of the control unit in pair j), and then calculate an average of those K differences, $\bar{d} = \frac{1}{K}\sum_{k=1}^{K} d_j$. This approach uses up many degrees of freedom. There is a growing literature devoted to optimal estimation strategies that can appropriately condition on pair status without losing too much power (see references section at the end of the chapter).

We will revisit the more general version of ignorability characterized in (16.2) in Chapter 18 because the analysis strategies in that chapter rely on this assumption. Thus another reason for having a strong conceptual understanding of the properties of randomized block experiments is that this will help us to understand the similar assumptions that are often made when adjusting for differences between treatment and control groups in observational studies, in which case there is no randomization to ensure that the model holds, hence the value in being explicit about assumptions.

Efficiency

Another design property was revealed in our discussion above. The more similar the potential outcomes are across treatment groups, the *closer* our estimate will be on average to the value of the estimand. Said another way, the more similar the units being compared across treatment groups, the smaller the variance of the randomization or sampling distribution.

Randomized block experiments. Ideally, the randomized block experiment creates subgroups (blocks) within which the y_i^0's are more similar to each other and the y_i^1's are more similar to each other across treatment groups. This makes it possible to get sharper estimates of block-specific treatment effects which can then be combined using a weighted average into an estimate of the average treatment effect across the entire experimental sample or population. A linear regression using block indicators as predictors achieves a similar result. This block-specific homogeneity is why the randomized block experiment yields estimates that have smaller standard errors on average than estimates from completely randomized experiments of the same sample size.

Increasing precision by adjusting for pre-treatment variables. Another strategy for achieving efficiency gains in treatment effect estimates using data from a randomized experiment is to adjust for pre-treatment variables that are predictive of the outcome. This requires regression modeling, but models that are fit to data from a randomized experiment tend to be robust to deviations from the parametric assumptions inherent in linear regression. (See Section 17.6 for a discussion of why you should not control in the regression for variables that occur *after* the treatment.)

Consider the data from the idealized randomized experiment displayed in Figure 16.3. If we estimate the treatment effect for these data using a regression only on the treatment indicator (so, effectively, a difference in means estimate) the standard error is 8.8. However if we include sex and age as predictors the standard error drops to 1.2! What's going on?

This randomization (that is, this random draw from the randomization distribution) in particular yielded exquisite balance, but under the completely randomized design the balance can vary simply by chance. The standard error in the model without additional predictors reflects the fact that we might not be so lucky the next time and thus our estimate might end up much farther from the truth. If, however, this balance had resulted from a randomized block or matched pairs experiment, then we would not expect the treatment effect estimates to vary nearly so much across repeated randomizations. Exercise 16.4 illustrates this point via simulation. Using a model that accounts for this design by conditioning on the blocking variables in our model would allow us to capture these efficiency gains.

When we condition on pre-treatment variables in the absence of such an intentional design we are capitalizing on the association between the potential outcomes and these variables to reduce the variance of our treatment effect estimate. With a pre-treatment variable that is very predictive of the response, this association creates homogeneity with regard to potential outcomes for observations with similar values of that pre-treatment variable. It is as if nature created a randomized block experiment and we are taking advantage of it.

Stable unit treatment value assumption (SUTVA): no interference across units and no hidden versions of the treatment

Our definition of potential outcomes from the start of this chapter implicitly encapsulates yet another assumption! Each person's potential outcome is defined in terms of only his or her own treatment assignment, $y_i^{z_i}$. One could imagine, alternatively, defining each potential outcome in terms of the *collection* of treatment assignments, z, across all study participants, y_i^z. In other words, person i's outcome would be a function not only of her own treatment assignment, but also the treatment assignments of others in the sample. This might make conceptual sense in situations where the outcome of one person could potentially be affected

by the treatment of others. This level of generality would quickly become intractable however. Even with our small study with two levels of the treatment, there are $2^8 = 256$ different possible allocations of treatments to units. We clearly do not have enough data to inform 256 potential outcomes for each person. For this reason, researchers often simply hope that such interference among units, or spillover, does not exist (for more recent advances and alternatives see the references at the end of the chapter).

We can formalize this *stable unit treatment value assumption* (SUTVA) by a statement such as,

$$\text{SUTVA: } \; y_i^z = y_i^{z'} \text{ if } z_i = z_i',$$

where z_i and z_i' indicate the i^{th} element of the vectors z and z', respectively. Thus if SUTVA holds, $y_i^z = y_i^{z_i}$, and we are back where we started with the number of potential outcomes equal to the number of treatments.

SUTVA also implies that there are no hidden versions of treatments. That is, all of the units receive the same well-defined treatment. If this requirement does not hold then the effect of z is not well defined because, in effect, z_i doesn't mean the same thing as z_j. Typically researchers focus on the no-interference aspect of SUTVA, perhaps reasoning that the definition of the treatment can always be broadened sufficiently to reflect various versions of treatments. However, this can be a dangerous path leading eventually to treatment effects that, even if unbiased, are virtually uninterpretable. It is best to try to ensure that treatment definitions and implementations are as homogeneous as possible across experimental units.

Examples of potential SUTVA violations abound. An experiment testing the effect of a new fertilizer by randomly assigning adjacent plots to treatment or control is a classic example. Fertilizer from one plot might leach into an adjacent plot assigned to receive no fertilizer and thus affect the yield in that control plot. Vaccines that reduce the probability of a contagious disease within a school, business, or community could easily violate SUTVA if the vaccine is actually effective. Consider an experiment that recruited families from the same public housing complex and randomized them to receive a voucher to move to a better neighborhood or not. This could suffer from interference if a given family moving might influence (positively or negatively) the well-being of another family that for instance was randomized to not receive the voucher.

Tensions may exist between what is optimal for an intervention to succeed and what is optimal to evaluate the magnitude of its impact. A program developer might reasonably *hope* for interference among units to increase the potential impact of a treatment. For instance, consider a behavioral intervention whose goal is to reduce bullying among students. If it were possible to introduce the intervention to a just a small fraction of the student body of the school and yet, by some diffusion mechanism, create impacts on all the students, that would be a positive social externality. However if we tried to estimate the magnitude of the effect of just such an intervention with an experiment that randomized at the student level, it would be virtually impossible to interpret the estimated effects.

If we are planning a study such as this in which SUTVA is not likely to hold due to interference, one option is to assign the treatment at the level of a group beyond which interference is not likely to occur. For instance, there is a long tradition in education research of assigning treatments at the classroom or school-level and then randomizing at this level of aggregation. In this way the students that might be most likely to transfer knowledge or behaviors amongst themselves are all receiving the same treatment. With this type of design the most conservative approach is to interpret results only at the group level. However, more sophisticated approaches now exist that attempt to directly model the impact of this interference or diffusion (see references at the end of the chapter).

External validity: Difficulty of extrapolating to situations and individuals outside the experimental context

Randomized experiments are widely regarded as ideal research design for identifying causal effects. The ability of this design to recover causal effects that pertain to the sample studied is often referred to as *internal validity*. One of the greatest limitations of randomized experiments, however, is the study sample comprises only those individuals who are are part of the study. Those who agree to participate in a study may be quite different than the population we are more interested in learning about. In addition, the effects being measured in an experimental setting may differ "in the wild." To the extent that the units in the study are not representative to the population of interest, and that the environment in the study differs in important ways from the outside world, we say that the design lacks *external validity*. But proponents of controlled experiments have pointed out that external validity is an issue with non-experimental studies as well. Moreover, there are statistical approaches that can help to generalize treatment effect estimates from randomized trials to different populations. These approaches will be discussed in the next chapter.

How the experiment affects the participants

Randomized experiments pose other challenges or "threats to validity," a number of which revolve around the interaction between the study design and patient or researcher behavior.

For instance, it is possible that simply participating in a study will change one's behavior, a phenomenon that has been called the "Hawthorne effect," in reference to a series of experiments conducted in the Western Electric's Hawthorne Works plant in the 1920's. The studies, designed to estimate the effect of light on productivity, purportedly revealed that workers in both experimental groups increased their productivity not based on their exposure to light but rather simply because they knew they were being observed. This interpretation of these data has since been challenged but the potential for this sort of effect surely still exists. This type of participant reaction can prevent the researcher from estimating the effect of the treatment that would have occurred naturally. As an example, in the set of hypothetical studies just described, if the researcher had decided to collect daily food diaries for each participant during the intervention year, such scrutiny might have led participants to alter their diets in ways that would affect subsequent blood pressure measurements. Relatedly, one may wonder if it is possible for participants to maintain their current diet once they have been alerted to the potential relationship between omega-3 and blood pressure (among other health issues). Members of the control group might (either consciously or inadvertently) start to include more fish in their diet, for example, thus possibly attenuating the effects of the supplements by altering the intended counterfactual condition. These types of expectancy effects have been documented not only in human studies but in studies with animal participants (even rats!).

However it is not only the patients' expectations and psychology around receiving the treatment and participating in the study that could be problematic. The professionals administering the treatment or intervention (doctors, teachers, social workers) or the researchers gathering the outcome data might also be influenced by knowing which treatment the study participant received. If so, they could inadvertently bias the results, particularly if the measurement of the outcome is subjective. For instance an assessor who knows that a given study participant has been taking an anti-depressant medication for two months might be more likely to rate his depression level as less severe than in the absence of such knowledge.

This sort of problem has led medical researchers (in particular) to favor *double-blind* research designs in which neither the patients nor the researchers (and in some cases even the professionals administering the treatment) have knowledge regarding which people are

receiving which treatment. This strategy won't work in all settings (subjects are bound to know whether or not they received a six-week job training intervention!) and doesn't cure all problems but can address some concerns. These designs also lend credibility to estimates of complier effects using instruments, as we discuss in Sections 19.1–19.2.

Missing data and noncompliance

A different set of concerns that are also present in observational studies are those of missing data and noncompliance with treatment assignment. Since pre-treatment variables are independent of the treatment assignment, missing pre-treatment variables can be ignored without affecting the internal validity. A simple difference in means can be used on the full sample. Or we can run a conditional analysis on the complete cases sample that deletes observations missing values for the predictors in the model and still achieve an unbiased treatment effect estimate for this subsample. On the other hand, either of these strategies could increase standard errors and the latter might focus inference on a less interesting segment of the analysis sample.

Missing outcome data are common, however, and pose greater challenges than missing pre-treatment data. That's because they are often more likely to occur for those in the control group than those in the treatment group because those in the treatment group are more likely to be emotionally invested in the study (and thus will be more likely to show up for the test or assessment or to fill out the required survey). Even if the missingness rates are the same across groups, if those participants for whom we are missing outcome data in the control group are different from those for whom we are missing outcome data in the treatment group (that is, the distributions of their potential outcomes are different), and the missing outcome data are ignored (that is, a complete case analysis is used), then the benefits of the initial randomization will be destroyed. If analyses are performed on the complete case sample then the resulting treatment effect estimates will be biased. We discuss this issue more thoroughly in Chapters 18–19.

Noncompliance with treatment assignment describes a situation in which, for instance, subjects choose not to take their medication or to participate in an intervention to which they were randomly assigned. In our hypothetical omega-3 fatty acids study, the onus was on the participants to remember to take the supplements as well as to continue to choose to keep taking them. Furthermore participants assigned to maintain their usual diet could have decided to start taking the supplements anyway. Human beings are capricious and any participant at any time could have failed to comply with their treatment assignment. This sort of noncompliance creates a disconnect between the effect of assignment and the effect of actually taking the treatment which makes it more challenging to estimate the latter effect. We discuss this issue and a potential solution in more detail in Section 19.2.

16.7 Bibliographic note

You can get a sense of different perspectives on causal inference from the recent textbooks by Angrist and Pischke (2008), Kennedy (2008), Imbens and Rubin (2015), Morgan and Winship (2014), Vanderweele (2015), and Hernan and Robins (2018).

The fundamental problem of causal inference and the potential outcome notation were formally introduced by Rubin (1974, 1978) who also coined the term ignorability. Related earlier work includes Neyman (1923) and Cox (1958). A helpful description of several common causal estimands can be found in Imai et al. (2008). Abadie and Imbens (2006) discuss conditional average treatment effects (CATE).

Campbell and Stanley (1963) is an early presentation of causal inference in experiments and observational studies from a social science perspective; see also Achen (1986). Shadish, Cook, and Campbell (2002), as well as earlier incarnations of that text, provide accessible

and intuitive explanations for threats to validity in causal inference, including interactions between study design and patient or researcher behavior.

The stable unit treatment value assumption (SUTVA) was defined by Rubin (1978); see also Sobel (2006) for a more recent discussion of the implications of SUTVA violations in the context of a public policy intervention and evaluation. Ainsley, Dyke, and Jenkyn (1995) and Besag and Higdon (1999) discuss spatial models for interference between units in agricultural experiments. Recent developments in modeling interference are described in Hong and Raudenbush (2006), Rosenbaum (2007), Hudgens and Halloran (2008), and Aronow and Samii (2012). Gelman (2004d) discusses treatment interactions in before-after studies.

Several different strategies have been proposed for implementing paired designs in the context of individual-level experimental designs with many pre-treatment variables; examples include Morris (1979), Hill et al. (2000), and Greevy et al. (2004). See Imai (2008) for a discussion of analysis of paired experiments. Pair matching in the context of cluster-randomized experiments has been discussed recently by several authors including King et al. (2009) and Zhang et al. (2011). See Murray et al. (2006), Gelman and Hill (2007), Middleton (2012) and Hill (2012), among others, for discussions of group-randomized experiments and causal analysis with multilevel data.

A classic reference on placebo effects is Beecher (1955); a more recent review can be found in Bingel et al. (2011). While commonly used, it turns out the phrase "Hawthorne effect" does not have a consistent and universal definition, as exemplified by a survey of the literature conducted by XXXX (2004). For an intriguing reanalysis of the studies that gave rise to the name Hawthorne effect see Levitt et al. (2011).

Dawid (2000) offers another perspective on the potential-outcome framework.

16.8 Exercises

1. Suppose you are interested in the effect of the presence of vending machines in schools on childhood obesity. What randomized experiment would you want to do (in a perfect world) to evaluate this question?
2. Suppose you are interested in the effect of smoking on lung cancer. What randomized experiment could you plausibly perform (in the real world) to evaluate this effect?
3. The table below describes a hypothetical experiment to evaluate the effect of a job training program on the hourly wages of 2000 people. Each row of the table specifies a category of person, as defined by his or her sex x, treatment indicator z, and potential outcomes y^0, y^1. For simplicity, we assume unrealistically that all the people in this experiment fit into these eight categories.

Category	# people in category	x	z	y^0	y^1
1	300	0	0	4	6
5	300	0	0	10	12
2	300	1	0	4	6
6	100	1	0	10	12
3	100	0	1	4	6
7	100	0	1	10	12
4	600	1	1	4	6
8	200	1	1	10	12

   In making the table we are assuming omniscience, so that we know both y^0 and y^1 for all observations. But the (nonomniscient) investigator would only observe x, T, and y^T for each unit. (For example, a person in category 1 would have $x=0, T=0, y=4$, and a person in category 3 would have $x=0, T=1, y=6$.)

   (a) What is the average treatment effect in this population of 2400 people?

(b) Is it plausible to believe that these data came from a randomized experiment? Defend your answer.

(c) Another population quantity is the mean of y for those who received the treatment minus the mea your estimate be of the average treatment effect if you were told that these data arose from an experiment that randomized separately by sex.

4. Use the covariate and potential outcome data from Figure 16.1 as a starting point for considering the randomization distributions of four different designs by creating simulations in R. Comment on the relative bias and efficiency of the four estimates listed below for each of the following designs:

(a) Completely randomized design;

(b) Randomized design blocked by the oldest four participants versus the youngest four;

(c) Matched pairs design.

Consider these estimates:

(a) Difference in means;

(b) Regression on treatment indicator and age;

(c) Regression on treatment indicator, age, and sex;

(d) Regression on treatment indicator, age, sex, and the interaction between the treatment and sex.

In addition, you can consider the matched pairs estimate the matched pairs experiment.

5. Write the functional form of the regression model that would yield the data in Figure 16.3.

6. Calculate how unlikely it is that you would see the level of imbalance displayed in Figure 16.4 or worse in the context of a completely randomized experiment. To do so you will first have to create a reasonable definition of imbalance. You can calculate this probability using mathematical computations or a simulation in R.

7. Show how the law of total probability (or law of iterated expectation) allows us to combine conditional estimates to get a marginal estimate.

8. The folder `Sesame` contains data from an experiment in which a randomly selected group of children was encouraged to watch the television program Sesame Street and the randomly selected control group was not.

(a) The goal of the experiment was to estimate the effect on child cognitive development of watching more Sesame Street. In the experiment, encouragement but not actual watching was randomized. Briefly explain why you think this was done. (Hint: think of practical as well as statistical reasons.)

(b) Suppose that the investigators instead had decided to test the effectiveness of the program simply by examining how test scores changed from before the intervention to after. What assumption would be required for this to be an appropriate causal inference? Use data on just the control group from this study to examine how realistic this assumption would have been.

9. Return to the Sesame Street example from the previous exercise. Did encouragement (the variable `viewenc` in the data) lead to an increase in post-test scores for letters (`postlet`) and numbers (`postnumb`)? Fit an appropriate model to answer this question.