Morgan / Stephen L., Winship / Christopher (2007): Counterfactuals and Causal Inference. Methods and Principles for Social Research. Cambridge University Press. ISBN: 978-0-521-67193-4

Chapter 2

The Counterfactual Model

In this chapter, we introduce the foundational components of the counterfactual model of causality, which is also known as the potential outcome model. We first discuss causal states and the relationship between potential and observed outcome variables. Then we introduce average causal effects and discuss the assumption of causal effect stability, which is maintained in most applications of the counterfactual model. We conclude with a discussion of simple estimation techniques, in which we demonstrate the importance of considering the relationship between the potential outcomes and the process of causal exposure.

2.1 Causal States and Potential Outcomes

For a binary cause, the counterfactual framework presupposes the existence of two well-defined causal states to which all members of the population of interest could be exposed.[1] These two states are usually labeled treatment and control. When a many-valued cause is analyzed, the convention is to refer to the alternative states as alternative treatments.

Consider the examples introduced in Section 1.3. Some of these examples have well-defined states, and others do not. The manpower training example is completely straightforward, and the two states are whether an individual is enrolled in a training program or not. The Catholic school example is similar. Here, the alternative states are "Catholic school" and "public school." The only complication with these examples is the possibility of inherent differences across training programs and Catholic schools. If any such treatment-site heterogeneity exists, then stratified analyses may be necessary, perhaps by regions

[1] We justify the importance of carefully defining the boundaries of the population of interest when presenting average causal effects later in this chapter. As we note there, we also provide an appendix to this chapter, in which we explain the general superpopulation model that we will adopt when the boundaries of the population can be clearly defined and when we have the good fortune of having a large random sample from the population.

of the country, size of the program, or whatever other dimension suggests that variability of the causal states deserves explicit modeling.[2]

Other examples have less clearly defined causal states. Consider the classic political participation line of inquiry. For the relationship between socioeconomic status and political participation, there are many underlying well-defined causal effects, such as the effect of having obtained at least a college degree on the frequency of voting in local elections and the effect of having a family income greater than some cutoff value on the amount of money donated to political campaigns. Well-defined causal states exist for these narrow causal effects, but it is not clear at all that well-defined causal states exist for the broad and internally differentiated concepts of socioeconomic status and political participation.

Finally, consider a related political science example. Beyond the voting technology effect discussed in Subsection 1.3.2 on the outcome of the 2000 presidential election, a broader set of question has been asked. To what extent do restrictions on who can vote determine who wins elections? A recent and highly publicized variant of this question is this: What is the effect on election outcomes of laws that forbid individuals with felony convictions from voting?[3] Uggen and Manza (2002) make the straightforward claim that the 2000 presidential election would have gone in favor of Al Gore if felons and ex-felons had been permitted to vote:

> Although the outcome of the extraordinarily close 2000 presidential election could have been altered by a large number of factors, it would almost certainly have been reversed had voting rights been extended to any category of disenfranchised felons. (Uggen and Manza 2002:792)

Uggen and Manza (2002) then note an important limitation of their conclusion:

> ... our counterfactual examples rely upon a ceteris paribus assumption – that nothing else about the candidates or election would change save the voting rights of felons and ex-felons. (Uggen and Manza 2002:795)

When thinking about this important qualification, one might surmise that a possible world in which felons had the right to vote would probably also be a world in which the issues (and probably candidates) of the election would be very different. Thus, the most challenging definitional issue here is not who counts as a felon or whether or not an individual is disenfranchised, but rather how well the alternative causal states can be characterized.

As this example illustrates, it is important that the "what would have been" nature of the conditionals that define the causal states of interest be carefully

[2]Hong and Raudenbush (2006) provide a careful analysis of retention policies in U.S. primary education, implementing this type of treatment-site stratification based on the average level of retention in different schools.

[3]Behrens, Uggen, and Manza (2003), Manza and Uggen (2004), and Uggen, Behrens, and Manza (2005) give historical perspective on this question.

laid out. When a *ceteris paribus* assumption is relied on to rule out other contrasts that are nearly certain to occur at the same time, the posited causal states are open to the charge that they are too metaphysical to justify the pursuit of causal analysis.[4]

Given the existence of well-defined causal states, causal inference in the counterfactual tradition proceeds by stipulating the existence of potential outcome random variables that are defined over all individuals in the population of interest. For a binary cause, we will denote potential outcome random variables as Y^1 and Y^0.[5] We will also adopt the notational convention from statistics in which realized values for random variables are denoted by lowercase letters. Accordingly, y_i^1 is the potential outcome in the treatment state for individual i, and y_i^0 is the potential outcome in the control state for individual i. The individual-level causal effect of the treatment is then defined as

$$\delta_i = y_i^1 - y_i^0. \tag{2.1}$$

Individual-level causal effects can be defined in ways other than as a linear difference in the potential outcomes. For example, the individual-level causal effect could be defined instead as the ratio of one individual-level potential outcome to another, as in y_i^1/y_i^0. In some applications, there may be advantages to these sorts of alternative definitions at the individual level, but the overwhelming majority of the literature represents individual-level causal effects as linear differences, as in Equation (2.1).[6]

[4]This may well be the case with the felon disenfranchisement example, but this is a matter for scholars in political sociology and criminology to debate. Even if the charge sticks, this particular line of research is nonetheless still an important contribution to the empirical literature on how changing laws to allow felons and ex-felons to vote could have potential effects on election outcomes.

[5]There is a wide variety of notation in the potential outcome and counterfactuals literature, and we have adopted the notation that we feel is the easiest to grasp. However, we should note that Equation (2.1) and its elements are often written as one of the following alternatives:

$$\begin{aligned} \Delta_i &= Y_{1i} - Y_{0i}, \\ \delta_i &= Y_i^t - Y_i^c, \\ \tau_i &= y_i(1) - y_i(0), \end{aligned}$$

and variants thereof. We use the right-hand superscript to denote the potential treatment state of the corresponding potential outcome variable, but other authors use the right-hand subscript or parenthetical notation. We also use numerical values to refer to the treatment states, but other authors (including us, see Morgan 2001, Winship and Morgan 1999, and Winship and Sobel 2004) use values such as t and c for the treatment and control states, respectively. There is also variation in the usage of uppercase and lowercase letters. We do not claim that everyone will agree that our notation is the easiest to grasp, and it is certainly not as general as, for example, the parenthetic notation. But it does seem to have proven itself in our own classes, offering the right balance between specificity and compactness.

[6]Moreover, the individual-level causal effect could be defined as the difference between the expectations of individual-specific random variables, as in $E[Y_i^1] - E[Y_i^0]$, where $E[.]$ is the expectation operator from probability theory (see, for a clear example of this alternative setup, King et al. 1994:76-82). In thinking about individuals self-selecting into alternative treatment states, it can be useful to set up the treatment effects in this way. In many applications, individuals are thought to consider potential outcomes with some recognition of

2.2 Treatment Groups and Observed Outcomes

For a binary cause with two causal states and associated potential outcome variables Y^1 and Y^0, the convention in the counterfactuals literature is to define a causal exposure variable, D, which takes on two values: D is equal to 1 for members of the population who are exposed to the treatment state and equal to 0 for members of the population who are exposed to the control state. Exposure to the alternative causal states is determined by a particular process, typically an individual's decision to enter one state or another, an outside actor's decision to allocate individuals to one state or another, a planned random allocation carried out by an investigator, or some combination of these alternatives.

By convention, those who are exposed to the treatment state are referred to as the treatment group whereas those who are exposed to the control state are referred to as the control group. Because D is defined as a population-level random variable (at least in most cases in observational data analysis), the treatment group and control group exist in the population as well as the observed data. Throughout this book, we will use this standard terminology, referring to treatment and control groups when discussing those who are exposed to alternative states of a binary cause. If more than two causal states are of interest, then we will shift to the semantics of alternative treatments and corresponding treatment groups, thereby discarding the baseline labels of control state and control group. Despite our adoption of this convention, we could rewrite all that follows referring to members of the population as what they are: those who are exposed to alternative causal states.

When we refer to individuals in the observed treatment and control groups, we will again adopt the notational convention from statistics in which realized values for random variables are denoted by lowercase letters. Accordingly, the random variable D takes on values of $d_i = 1$ for each individual i who is an observed member of the treatment group and $d_i = 0$ for each individual i who is an observed member of the control group.

Given these definitions of Y^1, Y^0, and D (as well as their realizations y_i^1, y_i^0, d_i), we can now define the observed outcome variable Y in terms of them. We can observe values for a variable Y as $y_i = y_i^1$ for individuals with $d_i = 1$ and as $y_i = y_i^0$ for individuals with $d_i = 0$. The observable outcome variable Y is therefore defined as

$$\begin{aligned} Y &= Y^1 \text{ if } D = 1, \\ Y &= Y^0 \text{ if } D = 0. \end{aligned}$$

inherent uncertainty of their beliefs, which may properly reflect true variability in their potential outcomes. But, when data for which a potential outcome is necessarily observed for any individual as a scalar value (via an observed outcome variable, defined later) are analyzed, this individual-level, random-variable definition is largely redundant. Accordingly, we will denote individual-level potential outcomes as values such as y_i^1 and y_i^0, regarding these as realizations of population-level random variables Y^1 and Y^0 while recognizing, at least implicitly, that they could also be regarded as realizations of individual-specific random variables Y_i^1 and Y_i^0.

Table 2.1: The Fundamental Problem of Causal Inference

Group	Y^1	Y^0
Treatment group $(D = 1)$	Observable as Y	Counterfactual
Control group $(D = 0)$	Counterfactual	Observable as Y

This paired definition is often written compactly as

$$Y = DY^1 + (1 - D)Y^0. \tag{2.2}$$

In words, one can never observe the potential outcome under the treatment state for those observed in the control state, and one can never observe the potential outcome under the control state for those observed in the treatment state. This impossibility implies that one can never calculate individual-level causal effects.

Holland (1986) describes this challenge as the fundamental problem of causal inference in his widely read introduction to the counterfactual model. Table 2.1 depicts the problem. Causal effects are defined within rows, which refer to groups of individuals in the treatment state or in the control state. However, only the diagonal of the table is observable, thereby rendering impossible the direct calculation of individual-level causal effects merely by means of observation and then subtraction.[7]

As shown clearly in Equation (2.2), the outcome variable Y, even if we could enumerate all of its individual-level values y_i in the population, reveals only half of the information contained in the underlying potential outcome variables. Individuals contribute outcome information only from the treatment state in which they are observed. This is another way of thinking about Holland's fundamental problem of causal inference. The outcome variables we must analyze – labor market earnings, test scores, and so on – contain only a portion of the information that would allow us to directly calculate causal effects for all individuals.

2.3 The Average Treatment Effect

Because it is typically impossible to calculate individual-level causal effects, we focus attention on the estimation of aggregated causal effects, usually alternative

[7] As Table 2.1 shows, we are more comfortable than some writers in using the label "counterfactual" when discussing potential outcomes. Rubin (2005), for example, avoids the term counterfactual, under the argument that potential outcomes become counterfactual only after treatment assignment has occurred. Thus no potential outcome is ever *ex ante* counterfactual. We agree, of course. But, because our focus is on observational data analysis, we find the counterfactual label useful for characterizing potential outcomes that are rendered unobservable *ex post* to the treatment assignment/selection mechanism.

average causal effects. With $E[.]$ denoting the expectation operator from probability theory, the average treatment effect in the population is

$$\begin{aligned} E[\delta] &= E[Y^1 - Y^0] \\ &= E[Y^1] - E[Y^0]. \end{aligned} \tag{2.3}$$

The second line of Equation (2.3) follows from the linearity of the expectation operator: The expectation of a difference is equal to the difference of the two expectations.[8]

For Equation (2.3), the expectation is defined with reference to the population of interest. For the political science examples in Chapter 1, the population could be "all eligible voters" or "all eligible voters in Florida." For other examples, such as the manpower training example, the population would be defined similarly as "all adults eligible for training," and eligibility would need to be defined carefully. Thus, to define average causal effects and then interpret estimates of them, it is crucial that researchers clearly define the characteristics of the individuals in the assumed population of interest.[9]

Note also that the subscripting on i for δ_i has been dropped for Equation (2.3). Even so, δ is not necessarily constant in the population, as it is a random variable just like Y^1 and Y^0. We can drop the subscript i in this equation because the expected causal effect of a randomly selected individual from the population is equal to the average causal effect across individuals in the population. We will at times throughout this book reintroduce redundant subscripting on i in order to reinforce the inherent individual-level heterogeneity of the potential outcomes and the causal effects they define, but we will be clear when we are doing so.

Consider the Catholic school example from Subsection 1.3.2 that demonstrates the relationship between observed and potential outcomes and how these are related to typical estimation of the average causal effect in Equation (2.3). For the Catholic school effect on learning, the potential outcome under the treatment, y_i^1, is the what-if achievement outcome of individual i if he or she were enrolled in a Catholic school. The potential outcome under the control, y_i^0, is the what-if achievement outcome of individual i if he or she were enrolled in a public school. Accordingly, the individual-level causal effect, δ_i, is the what-if difference in achievement that could be calculated if we could simultaneously educate individual i in both a Catholic school and a public school.[10] The average

[8] However, more deeply, it also follows from the assumption that the causal effect is defined as a linear difference at the individual level, which allows the application of expectations in this simple way to characterize population-level average effects.

[9] And, regardless of the characterization of the features of the population, we will assume throughout this book that the population is a realization of an infinite superpopulation. We discuss our decision to adopt this underlying population model in an appendix to this chapter. Although not essential to understanding most of the material in this book, some readers may find it helpful to read that appendix now in order to understand how these definitional issues are typically settled in this literature.

[10] However, it is a bit more complex than this. Now that we have introduced a real-world scenario, other assumptions must also be invoked, notably the stable unit treatment value assumption, introduced and explained in the next section.

causal effect, $E[\delta]$, is then the mean value among all students in the population of these what-if differences in test scores. The average causal effect is also equal to the expected value of the what-if difference in test scores for a randomly selected student from the population.

2.4 The Stable Unit Treatment Value Assumption

In most applications, the counterfactual model retains its transparency through the maintenance of a very simple but strong assumption known as the stable unit treatment value assumption or SUTVA (see Rubin 1980b, 1986). In economics, this is sometimes referred to as a no-macro-effect or partial equilibrium assumption (see Heckman 2000, 2005 and Garfinkel, Manski, and Michalopoulos 1992 for the history of these ideas and Manski and Garfinkel 1992 for examples). SUTVA, as implied by its name, is a basic assumption of causal effect stability that requires that the potential outcomes of individuals be unaffected by potential changes in the treatment exposures of other individuals. In the words of Rubin (1986:961), who developed the term,

> SUTVA is simply the a priori assumption that the value of Y for unit u when exposed to treatment t will be the same no matter what mechanism is used to assign treatment t to unit u and no matter what treatments the other units receive.

Consider the idealized example in Table 2.2, in which SUTVA is violated because the treatment effect varies with treatment assignment patterns. For the idealized example, there are three randomly drawn subjects from a population of interest, and the study is designed such that at least one of the three study subjects must receive the treatment and at least one must receive the control. The first column of the table gives the six possible treatment assignment patterns.The first row of Table 2.2 presents all three ways to assign one individual to the treatment and the other two to the control, as well as the potential outcomes for each of the three subjects. Subtraction within the last column shows that the individual-level causal effect is 2 for all three individuals. The second row of Table 2.2 presents all three ways to assign two individuals to the treatment and one to the control. As shown in the last column of the row, the individual-level causal effects implied by the potential outcomes are now 1 instead of 2. Thus, for this idealized example, the underlying causal effects are a function of the treatment assignment patterns, such that the treatment is less effective when more individuals are assigned to it. For SUTVA to hold, the potential outcomes would need to be identical for both rows of the table.

This type of treatment effect dilution is only one way in which SUTVA can be violated. More generally, suppose that $\mathbf{d}$ is an $N \times 1$ vector of treatment indicator variables for N individuals (analogous to the treatment assignment vectors in the first column of Table 2.2), and define potential outcomes generally as functions of the vector $\mathbf{d}$. The outcome for individual i under the

Table 2.2: A Hypothetical Example in Which SUTVA is Violated

Treatment assignment patterns			Potential outcomes	
$\begin{bmatrix} d_1 = 1 \\ d_2 = 0 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 1 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 0 \\ d_3 = 1 \end{bmatrix}$	$y_1^1 = 3$ $y_2^1 = 3$ $y_3^1 = 3$	$y_1^0 = 1$ $y_2^0 = 1$ $y_3^0 = 1$
$\begin{bmatrix} d_1 = 1 \\ d_2 = 1 \\ d_3 = 0 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 0 \\ d_2 = 1 \\ d_3 = 1 \end{bmatrix}$ or	$\begin{bmatrix} d_1 = 1 \\ d_2 = 0 \\ d_3 = 1 \end{bmatrix}$	$y_1^1 = 2$ $y_2^1 = 2$ $y_3^1 = 2$	$y_1^0 = 1$ $y_2^0 = 1$ $y_3^0 = 1$

treatment is $y_i^1(\mathbf{d})$, and the outcome for individual i under the control is $y_i^0(\mathbf{d})$. Accordingly, the individual-level causal effect for individual i is $\delta_i(\mathbf{d})$. SUTVA is what allows us to write $y_i^1 = y_i^1(\mathbf{d})$ and $y_i^0 = y_i^0(\mathbf{d})$ and, as a result, assert that individual-level causal effects δ_i exist that are independent of the assignment process itself.[11]

Sometimes it is argued that SUTVA is so restrictive that we need an alternative conception of causality for the social sciences. We agree that SUTVA is very sobering. However, our position is that SUTVA reveals the limitations of observational data and the perils of immodest causal modeling rather than the limitations of the counterfactual model itself. Rather than consider SUTVA as overly restrictive, researchers should always reflect on the plausibility of SUTVA in each application and use such reflection to motivate a clear discussion of the meaning and scope of a causal effect estimate.

Consider the example of the Catholic school effect again. For SUTVA to hold, the effectiveness of Catholic schooling cannot be a function of the number (and/or composition) of students who enter the Catholic school sector. For a variety of reasons – endogenous peer effects, capacity constraints, and so on – most school effects researchers would probably expect that the Catholic school effect would change if large numbers of public school students entered the Catholic school sector. As a result, because there are good theoretical reasons to believe that macro effects would emerge if Catholic school enrollments ballooned, it may be that researchers can estimate the causal effect of Catholic schooling only for those who would typically choose to attend Catholic schools, but also subject to the constraint that the proportion of students educated in Catholic schools remain relatively constant. Accordingly, it may be impossible to determine from any data that could be collected what the Catholic school effect on achievement would be under a new distribution of students across school sectors that would result from a large and effective policy intervention.

[11] In other words, if SUTVA is violated, then Equation (2.1) must be written in its most general form as $\delta_i(\mathbf{d}) = y_i^1(\mathbf{d}) - y_i^0(\mathbf{d})$. In this case, individual-level treatment effects could be different for every possible configuration of treatment exposures.

As a result, the implications of research on the Catholic school effect for research on school voucher programs (see Subsection 1.3.2) may be quite limited, and this has not been clearly enough recognized by some (see Howell and Peterson 2002, Chapter 6).

Consider also the manpower training example introduced in Subsection 1.3.2. Here, the suitability of SUTVA may depend on the particular training program. For small training programs situated in large labor markets, the structure of wage offers to retrained workers may be entirely unaffected by the existence of the training program. However, for a sizable training program in a small labor market, it is possible that the wages on offer to retrained workers would be a function of the way in which the price of labor in the local labor market responds to the movement of trainees in and out of the program (as might be the case in a small company town after the company has just gone out of business and a training program is established). As a result, SUTVA may be reasonable only for a subset of the training sites for which data have been collected.

Finally, consider SUTVA in the context of an example that we will not consider in much detail in this book: the evaluation of the effectiveness of mandatory school desegregation plans in the 1970s on the subsequent achievement of black students. Gathering together the results of a decade of research, Crain and Mahard (1983) conducted a meta-analysis of 93 studies of the desegregation effect on achievement. They argued that the evidence suggests an increase of .3 standard deviations in the test scores of black students across all studies.[12] It seems undeniable that SUTVA is violated for this example, as the effect of moving from one school to another must be a function of relative shifts in racial composition across schools. Breaking the analysis into subsets of cities where the compositional shifts were similar could yield average treatment effect estimates that can be more clearly interpreted. In this case, SUTVA would be abandoned in the collection of all desegregation events, but it could then be maintained for some groups (perhaps in cities where the compositional shift was relatively small).

In general, if SUTVA is maintained but there is some doubt about its validity, then certain types of marginal effect estimates can usually still be defended. The idea here would be to state that the estimates of average causal effects hold only for what-if movements of a very small number of individuals from one hypothetical treatment state to another. If more extensive what-if contrasts are of interest, such as would be induced by a widespread intervention, then SUTVA would need to be dropped and variation of the causal effect as a function of

[12] As reviewed by Schofield (1995) and noted in Clotfelter (2004), most scholars now accept that the evidence suggests that black students who were bused to predominantly white schools experienced small positive reading gains but no substantial mathematics gains. Cook and Evans (2000:792) conclude that "... it is unlikely that efforts at integrating schools have been an important part of the convergence in academic performance [between whites and blacks], at least since the early 1970s" (see also Armor 1995; Rossell, Armor, and Walberg 2002). Even so, others have argued that the focus on test score gains has obscured some of the true effectiveness of desegregation. In a review of these longer-term effects, Wells and Crain (1994:552) conclude that "interracial contact in elementary and secondary school can help blacks overcome perpetual segregation."

treatment assignment patterns would need to be modeled explicitly. This sort of modeling can be very challenging and generally requires a full model of causal effect exposure that is grounded on a believable theoretical model that sustains subtle predictions about alternative patterns of individual behavior. But it is not impossible, and it represents a frontier of research in many well-established causal controversies (see Heckman 2005, Sobel 2006).

2.5 Treatment Assignment and Observational Studies

A researcher who wishes to estimate the effect of a treatment that he or she can control on an outcome of interest typically designs an experiment in which subjects are randomly assigned to alternative treatment and control groups. Other types of experiments are possible, as we described earlier in Chapter 1, but randomized experiments are the most common research design when researchers have control over the assignment of the treatment.

After randomization of the treatment, the experiment is run and the values of the observed outcome, y_i, are recorded for those in the treatment group and for those in the control group. The mean difference in the observed outcomes across the two groups is then anointed the estimated average causal effect, and discussion (and any ensuing debate) then moves on to the particular features of the experimental protocol and the degree to which the pool of study participants reflects the population of interest for which one would wish to know the average treatment effect.

Consider this randomization research design with reference to the underlying potential outcomes defined earlier. For randomized experiments, the treatment indicator variable D is forced by design to be independent of the potential outcome variables Y^1 and Y^0. (However, for any single experiment with a finite set of subjects, the values of d_i will be related to the values of y_i^1 and y_i^0 because of chance variability.) Knowing whether or not a subject is assigned to the treatment group in a randomized experiment yields no information whatsoever about a subject's what-if outcome under the treatment state, y_i^1, or, equivalently, about a subject's what-if outcome under the control state, y_i^0. Treatment status is therefore independent of the potential outcomes, and the treatment assignment mechanism is said to be ignorable.[13] This independence assumption is usually written as

$$(Y^0, Y^1) \perp\!\!\!\perp D, \tag{2.4}$$

where the symbol $\perp\!\!\!\perp$ denotes independence and where the parentheses enclosing

[13]Ignorability holds in the weaker situation in which S is a set of observed variables that completely characterize treatment assignment patterns and in which $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$. Thus treatment assignment is ignorable when the potential outcomes are independent of D, conditional on S. We will offer a more complete discussion of ignorability in the next three chapters.

Y^0 and Y^1 stipulate that D must be jointly independent of all functions of the potential outcomes (such as δ). For a properly run randomized experiment, learning the treatment to which a subject has been exposed gives no information whatsoever about the size of the treatment effect.

At first exposure, this way of thinking about randomized experiments and potential outcomes can be confusing. The independence relationships represented by Equation (2.4) seem to imply that even a well-designed randomized experiment cannot tell us about the causal effect of the treatment on the outcome of interest. But, of course, this is not so, as Equation (2.4) does not imply that D is independent of Y. If individuals are randomly assigned to both the treatment and the control states, and individual causal effects are nonzero, then the definition of the outcome variable, $Y = DY^1 + (1-D)Y^0$ in Equation (2.2), ensures that Y and D will be dependent.

Now consider the additional challenges posed by observational data analysis. It is the challenges to causal inference that are the defining features of an observational study according to Rosenbaum (2002:vii):

> An *observational study* is an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects.

This definition is consistent with the Cox and Reid definition quoted in Chapter 1 (see page 7).

Observational data analysis in the counterfactual tradition is thus defined by a lack of control over the treatment [and, often more narrowly by the infeasibility of randomization designs that allow for the straightforward maintenance of the independence assumption in Equation (2.4)]. An observational researcher, hoping to estimate a causal effect, begins with observed data in the form of values $\{y_i, d_i\}_i^N$ for an observed outcome variable, Y, and a treatment status variable, D. To determine the causal of effect of D on Y, the first step in analysis is to investigate the treatment selection mechanism. Notice the switch in language from assignment to selection. Because observational data analysis is defined as empirical inquiry in which the researcher does not have the capacity to assign individuals to treatments (or, as Rosenbaum states equivalently, to assign treatments to individuals), researchers must instead investigate how individuals end up in alternative treatment states.

And herein lies the challenge of much scholarship in the social sciences. Although some of the process by which individuals select alternative treatments can be examined empirically, a full accounting of treatment selection is sometimes impossible (e.g., if subjects are motivated to select on the causal effect itself and a researcher does not have a valid measure of their expectations). As much as this challenge may be depressing to a dispassionate policy designer/evaluator, this predicament should not be depressing for social scientists in general. On the contrary, our existential justification rests on the pervasive need to deduce theoretically from a set of basic principles or infer from experience and knowledge of related studies the set of defendable assumptions about

the missing components of the treatment selection mechanism. Only through such effort can it be determined whether causal analysis can proceed or whether further data collection and preliminary theoretical analysis are necessary.

2.6 Average Causal Effects and Naive Estimation

As described in prior sections of this chapter, the fundamental problem of causal inference requires that we focus on non-individual-level causal effects, maintaining assumptions about treatment assignment and treatment stability that will allow us to give causal interpretations to differences in average values of observed outcomes. In the remainder of this chapter, we define average treatment effects of varying sorts and then lay out the complications of estimating them. In particular, we consider how average treatment effects vary across those who receive the treatment and those who do not.

2.6.1 Conditional Average Treatment Effects

The average causal effect, known as the average treatment effect in the counterfactual tradition, was defined in Equation (2.3) as $E[\delta] = E[Y^1 - Y^0]$. This average causal effect is the most common subject of investigation in the social sciences, and it is the causal effect that is closest to the sorts of effects investigated in the three broad foundational examples introduced in Chapter 1: the effects of family background and mental ability on educational attainment, the effects of educational attainment and mental ability on earnings, and the effects of socioeconomic status on political participation. More narrowly defined average causal effects are of interest as well in virtually all of the other examples introduced in Chapter 1.

Two conditional average treatment effects are of particular interest. The average treatment effect for those who typically take the treatment is

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1], \end{aligned} \tag{2.5}$$

and the average treatment effect for those who typically do not take the treatment is

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0], \end{aligned} \tag{2.6}$$

where, as for the average treatment effect in Equation (2.3), the second line of each definition follows from the linearity of the expectation operator. These two conditional average causal effects are often referred to by the acronyms ATT and ATC, which signify the average treatment effect for the treated and the average treatment effect for the controls, respectively.

Consider the examples again. For the Catholic school example, the average treatment effect for the treated is the average effect of Catholic schooling on the achievement of those who typically attend Catholic schools rather than across all students who could potentially attend Catholic schools. The difference between the average treatment effect and the average treatment effect for the treated can also be understood with reference to individuals. From this perspective, the average treatment effect in Equation (2.3) is the expected what-if difference in achievement that would be observed if we could educate a randomly selected student in both a public school and a Catholic school. In contrast, the average treatment effect for the treated in Equation (2.5) is the expected what-if difference in achievement that would be observed if we could educate a randomly selected Catholic school student in both a public school and a Catholic school.

For this example, the average treatment effect among the treated is a theoretically important quantity, for if there is no Catholic school effect for Catholic school students, then most reasonable theoretical arguments would maintain that it is unlikely that there would be a Catholic school effect for students who typically attend public schools (at least after adjustments for observable differences between Catholic and public school students). And, if policy interest were focused on whether or not Catholic schooling is beneficial for Catholic school students (and thus whether public support of transportation to Catholic schools is a benevolent government expenditure, etc.), then the Catholic school effect for Catholic school students is the only quantity we would want to estimate. The treatment effect for the untreated would be of interest as well if the goal of analysis is ultimately to determine the effect of a potential policy intervention, such as a new school voucher program, designed to move more students out of public schools and into Catholic schools. In fact, an even narrower conditional treatment effect might be of interest: $E[\delta|D = 0, \text{CurrentSchool} = \text{Failing}]$, where of course the definition of being currently educated in a failing school would have to be clearly specified.

The manpower training example is similar, in that the subject of first investigation is surely the treatment effect for the treated (as discussed in detail in Heckman et al. 1999). If a cost-benefit analysis of a program is desired, then a comparison of the aggregate net benefits for the treated to the overall costs of the program to the funders is needed. The treatment effect for other potential enrollees in the treatment program could be of interest as well, but this effect is secondary (and may be impossible to estimate for groups of individuals completely unlike those who have enrolled in the program in the past).

The butterfly ballot example is somewhat different as well. Here, the treatment effect of interest is bound by a narrow question that was shaped by media attention. The investigators were interested only in what actually happened in the 2000 election, and they focused very narrowly on whether the effect of having had a butterfly ballot rather than an optical scan ballot caused some individuals to miscast their votes. And, in fact, they were most interested in narrow subsets of the treated, for whom specific assumptions were more easily asserted and defended (e.g., those who voted for Democrats in all other races on the ballot but who voted for Pat Buchanan or Al Gore for president). In this

case, the treatment effect for the untreated, and hence the all-encompassing average treatment effect, was of little interest to the investigators (or to the contestants and the media).

As these examples demonstrate, more specific average causal effects (or more general properties of the distribution of causal effects) are often of greater interest than simply the average causal effect in the population. In this book, we will focus mostly on the three types of average causal effects represented by Equations (2.3), (2.5), and (2.6), as well as simple conditional variants of them. But, especially when presenting instrumental variable estimators later and discussing general heterogeneity issues, we will also focus on more narrowly defined causal effects. Heckman (2000), Manski (1995), and Rosenbaum (2002) all give full discussions of the variety of causal effects that may be relevant for different types of applications, such as quantiles of the distribution of individual-level causal effects in subpopulations of interest and the probability that the individual-level causal effect is greater than zero among the treated (see also Heckman, Smith, and Clements 1997).

2.6.2 Naive Estimation of Average Treatment Effects

Suppose again that randomization of the treatment is infeasible and thus that only an observational study is possible. Instead, an autonomous fixed treatment selection regime prevails, where π is the proportion of the population of interest that takes the treatment instead of the control. In this scenario, the value of π is fixed in the population by the behavior of individuals, and it is unknown. Suppose further that we have observed survey data from a relatively large random sample of the population of interest.

Because we are now shifting from the population to data generated from a random sample of the population, we must use appropriate notation to distinguish sample-based quantities from the population-based quantities that we have considered until now. For the sample expectation of a quantity in a sample of size N, we will use a subscript on the expectation operator, as in $E_N[.]$. With this notation, $E_N[d_i]$ is the sample mean of the dummy treatment variable, $E_N[y_i|d_i = 1]$ is the sample mean of the outcome for those observed in the treatment group, and $E_N[y_i|d_i = 0]$ is the sample mean of the outcome for those observed in the control group.[14] The naive estimator of the average causal effect is then defined as

$$\hat{\delta}_{\text{NAIVE}} \equiv E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0], \tag{2.7}$$

which is simply the difference in the sample means of the observed outcome variable Y for the observed treatment and control groups.

In observational studies, the naive estimator rarely yields a consistent estimate of the average treatment effect because it converges to a contrast,

[14] In other words, the subscript N serves the same basic notational function as an overbar on y_i, as in $\bar{y}_i$. We use this sub-N notation, as it allows for greater clarity in aligning sample and population-level conditional expectations for subsequent expressions.

$E[Y|D = 1] - E[Y|D = 0]$, that is not equivalent to (and usually not equal to) any of the average causal effects defined earlier. To see why, decompose the average treatment effect in Equation (2.3) as

$$\begin{aligned} E[\delta] = &\{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &- \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\}. \end{aligned} \tag{2.8}$$

The average treatment effect is then a function of five unknowns: the proportion of the population that is assigned to (or self-selects into) the treatment along with four conditional expectations of the potential outcomes. Without introducing additional assumptions, we can consistently estimate with observational data from a random sample of the population only three of the five unknowns on the right-hand side of Equation (2.8), as we now show.

We know that, for a very large random sample, the mean of realized values for the dummy treatment variable D would be equal to the true proportion of the population that would be assigned to (or would select into) the treatment. More precisely, we know that the sample mean of the values d_i converges in probability to π, which we write as

$$E_N[d_i] \xrightarrow{p} \pi. \tag{2.9}$$

Although the notation of Equation (2.9) may appear unfamiliar, the claim is that, as the sample size N increases, the sample mean of the values d_i approaches the true value of π, which we assume is a fixed population parameter equal exactly to $E[D]$. Thus, the notation $\xrightarrow{p}$ denotes convergence in probability for a sequence of estimates over a set of samples where the sample size N is increasing to infinity.[15] We can offer similar claims about two other unknowns in Equation (2.8):

$$E_N[y_i|d_i = 1] \xrightarrow{p} E[Y^1|D = 1], \tag{2.10}$$

$$E_N[y_i|d_i = 0] \xrightarrow{p} E[Y^0|D = 0], \tag{2.11}$$

which indicate that the sample mean of the observed outcome in the treatment group converges to the true average outcome under the treatment state for those in the treatment group (and analogously for the control group and control state).

Unfortunately, however, there is no assumption-free way to effectively estimate the two remaining unknowns in Equation (2.8): $E[Y^1|D = 0]$ and $E[Y^0|D = 1]$. These are counterfactual conditional expectations: the average outcome under the treatment for those in the control group and the average outcome under the control for those in the treatment group. Without further assumptions, no estimated quantity based on observed data from a random sample of the population of interest would converge to the true values for these unknown counterfactual conditional expectations. For the Catholic school example, these are the average achievement of public school students if they had instead been

[15] Again, see our appendix to this chapter on our assumed superpopulation model.

educated in Catholic schools and the average achievement of Catholic school students if they had instead been educated in public schools.

2.6.3 Expected Bias of the Naive Estimator

In the last subsection, we noted that the naive estimator $\hat{\delta}_{\text{NAIVE}}$, which is defined as $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$, converges to $E[Y^1|D = 1] - E[Y^0|D = 0]$. In this subsection, we show why this contrast can be uninformative about the causal effect of interest in an observational study by analyzing the expected bias in the naive estimator as an estimator of the average treatment effect.[16] Consider the following rearrangement of the decomposition in Equation (2.8):

$$\begin{aligned} E[Y^1|D = 1] - E[Y^0|D = 0] \quad &= \quad E[\delta] \\ &\quad + \{E[Y^0|D = 1] - E[Y^0|D = 0]\} \\ &\quad + (1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\}. \end{aligned} \tag{2.12}$$

The naive estimator converges to the left-hand side of this equation, and thus the right-hand side shows both the true average treatment effect, $E[\delta]$, plus the expectations of two potential sources of expected bias in the naive estimator.[17] The first source of potential bias, $\{E[Y^0|D = 1] - E[Y^0|D = 0]\}$, is a *baseline bias* equal to the difference in the average outcome in the absence of the treatment between those in the treatment group and those in the control group. The second source of potential bias, $(1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\}$, is a *differential treatment effect bias* equal to the expected difference in the treatment effect between those in the treatment and those in the control group (multiplied by the proportion of the population under the fixed treatment selection regime that does not select into the treatment).

To clarify this decomposition of the bias of the naive estimator, consider a substantive example – the effect of education on an individual's mental ability. Assume that the treatment is college attendance. After administering a test to a group of young adults, we find that individuals who have attended college score higher than individuals who have not attended college. There are three possible reasons that we might observe this finding. First, attending college might make individuals smarter on average. This effect is the average treatment effect, represented by $E[\delta]$ in Equation (2.12). Second, individuals who

[16] An important point of this literature is that the bias of an estimator is a function of what is being estimated. Because there are many causal effects that can be estimated, general statements about the bias of particular estimators are always conditional on a clear indication of the causal parameter of interest.

[17] The referenced rearrangement is simply a matter of algebra. Let $E[\delta] = e$, $E[Y^1|D = 1] = a$, $E[Y^1|D = 0] = b$, $E[Y^0|D = 1] = c$, and $E[Y^0|D = 0] = d$ so that Equation (2.8) can be written more compactly as $e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$. Rearranging this expression as $a - d = e + a - b - \pi a + \pi b + \pi c - \pi d$ then simplifies to $a - d = e + \{c - d\} + \{(1 - \pi)[(a - c) - (b - d)]\}$. Substituting for a, b, c, d, and e then yields Equation (2.12).

Table 2.3: An Example of the Bias of the Naive Estimator

Group	$E[Y^1\|.]$	$E[Y^0\|.]$
Treatment group ($D = 1$)	10	6
Control group ($D = 0$)	8	5

attend college might have been smarter in the first place. This source of bias is the baseline difference represented by $E[Y^0|D = 1] - E[Y^0|D = 0]$. Third, the mental ability of those who attend college may increase more than would the mental ability of those who did not attend college if they had instead attended college. This source of bias is the differential effect of the treatment, represented by $E[\delta|D = 1] - E[\delta|D = 0]$.

To further clarify the last term in the decomposition, consider the alternative hypothetical example depicted in Table 2.3. Suppose, for context, that the potential outcomes are now some form of labor market outcome, and that the treatment is whether or not an individual has obtained a college degree. Suppose further that 30 percent of the population obtains college degrees, such that π is equal to .3. As shown on the main diagonal of Table 2.3, the average (or expected) potential outcome under the treatment is 10 for those in the treatment group, and the average (or expected) potential outcome under the control for those in the control group is 5. Now, consider the off-diagonal elements of the table, which represent the counterfactual average potential outcomes. According to these values, those who have college degrees would have done better in the labor market than those without college degrees in the counterfactual state in which they did not in fact obtain college degrees (i.e., on average they would have received 6 instead of 5). Likewise, those who do not obtain college degrees would not have done as well as those who did obtain college degrees in the counterfactual state in which they did in fact obtain college degrees (i.e., on average they would have received only 8 instead of 10). Accordingly, the average treatment effect for the treated is 4, whereas the average treatment effect for the untreated is only 3. Finally, if the proportion of the population that completes college is .3, then the average treatment effect is 3.3, which is equal to $.3(10 - 6) + (1 - .3)(8 - 5)$.

Consider now the bias in the naive estimator. For this example, the naive estimator, as defined in Equation (2.7), would be equal to 5, on average, across repeated samples from the population (i.e., because $E[Y^1|D = 1] - E[Y^0|D = 0] = 10 - 5$). Thus, over repeated samples, the naive estimator would be upwardly biased for the average treatment effect (i.e., yielding 5 rather than 3.3), the average treatment effect for the treated (i.e., yielding 5 rather than 4), and the average treatment effect for the untreated (i.e., yielding 5 rather than 3). Equation (2.12) gives the components of the total expected bias of 1.7 for the naive estimator as an estimate of the average treatment effect. The term $\{E[Y^0|D = 1] - E[Y^0|D = 0]\}$, which we labeled the expected baseline bias, is

$6 - 5 = 1$. The term $(1 - \pi)\{E[\delta|D = 1] - E[\delta|D = 0]\}$, which is the expected differential treatment effect bias, is $(1 - .3)(4 - 3) = .7$.[18]

2.6.4 Estimating Causal Effects Under Maintained Assumptions About Potential Outcomes

What assumptions suffice to enable unbiased and consistent estimation of the average treatment effect with the naive estimator? There are two basic classes of assumptions: (1) assumptions about potential outcomes for subsets of the population defined by treatment status and (2) assumptions about the treatment assignment/selection process in relation to the potential outcomes. These two types of assumptions are variants of each other, and each may have a particular advantage in motivating analysis in a particular application.

In this section, we discuss only the first type of assumption, as it suffices for the present examination of the fallibility of the naive estimator. And our point in introducing these assumptions is simply to explain in one final way why the naive estimator will fail in most social science applications to generate an unbiased and consistent estimate of the average causal effect when randomization of the treatment is infeasible.

Consider the following two assumptions:

$$\text{Assumption 1: } E[Y^1|D = 1] = E[Y^1|D = 0], \tag{2.13}$$

$$\text{Assumption 2: } E[Y^0|D = 1] = E[Y^0|D = 0]. \tag{2.14}$$

If one asserts these two equalities and then substitutes into Equation (2.8), the number of unknowns is reduced from the original five parameters to the three parameters that we know from Equations (2.9)–(2.11) can be consistently estimated with data generated from a random sample of the population. If both Assumptions 1 and 2 are maintained, then the average treatment effect, the average treatment effect for the treated, and the average treatment effect for the untreated in Equations (2.3), (2.5), and (2.6), respectively, are all equal. And the naive estimator is consistent for all of them.

When would Assumptions 1 and 2 in Equations (2.13) and (2.14) be reasonable? Clearly, if the independence of potential outcomes, as expressed in Equation (2.4), is valid because the treatment has been randomly assigned, then Assumptions 1 and 2 in Equations (2.13) and (2.14) are implied. But, for observational data analysis, for which random assignment is infeasible, these assumptions would rarely be justified.

Consider the Catholic school example introduced in Subsection 1.3.2. If one were willing to assume that those who choose to attend Catholic schools

[18] In general, the amount of this expected differential treatment effect bias declines as more of the population is characterized by the treatment effect for the treated than by the treatment effect for the untreated (i.e., as π approaches 1).

do so for completely random reasons, then these two assumptions could be asserted. But we know from the applied literature that this characterization of treatment selection is false. Nonetheless, one might be able to assert instead a weaker narrative to warrant these two assumptions. One could maintain that students and their parents make enrollment decisions based on tastes for an education with a religious foundation and that this taste is unrelated to the two potential outcomes, such that those with a taste for the religious foundations of education would not necessarily benefit more from actually being educated in a Catholic school than in other schools. This possibility also seems unlikely, in part because it implies that those with a distaste for a religious education do not attend Catholic schools and it seems reasonable to assume that they would perform substantially worse in a Catholic school than the typical student who does attend a Catholic school.

Thus, at least for the Catholic school example, there seems no way to justify the naive estimator as an unbiased and consistent estimator of the average treatment effect (or of the average treatment effect for the treated and the average treatment effect for the untreated). We encourage the reader to consider all of the examples presented in the first chapter, and we suspect that all will agree that Assumptions 1 and 2 in Equations (2.13) and (2.14) cannot be sustained for any of them.

But it is important to recognize that assumptions such as these can (and should) be evaluated separately. Consider the two relevant cases for Assumptions 1 and 2:

1. If Assumption 1 is true but Assumption 2 is not, then $E[Y^1|D = 1] = E[Y^1|D = 0]$ whereas $E[Y^0|D = 1] \neq E[Y^0|D = 0]$. In this case, the naive estimator remains biased and inconsistent for the average treatment effect, but it is now unbiased and consistent for the average treatment effect for the untreated. This result is true because of the same sort of substitution we noted earlier. We know that the naive estimator $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$ converges to $E[Y^1|D = 1] - E[Y^0|D = 0]$. If Assumption 1 is true, then one can substitute $E[Y^1|D = 0]$ for $E[Y^1|D = 1]$. Then, one can state that the naive estimator converges to the contrast $E[Y^1|D = 0] - E[Y^0|D = 0]$ when Assumption 1 is true. This contrast is defined in Equation (2.6) as the average treatment effect for the untreated.

2. If Assumption 2 is true but Assumption 1 is not, then $E[Y^0|D = 1] = E[Y^0|D = 0]$ whereas $E[Y^1|D = 1] \neq E[Y^1|D = 0]$. The opposite result to the prior case follows. One can substitute $E[Y^0|D = 1]$ for $E[Y^0|D = 0]$ in the contrast $E[Y^1|D = 1] - E[Y^0|D = 0]$. Then, one can state that the naive estimator converges to the contrast $E[Y^1|D = 1] - E[Y^0|D = 1]$ when Assumption 2 is true. This contrast is defined in Equation (2.5) as the average treatment effect for the treated.

Considering the validity of Assumptions 1 and 2 separately shows that the naive estimator may be biased and inconsistent for the average treatment effect and yet may be unbiased and consistent for either the average treatment

effect for the treated or the average treatment effect for the untreated. These possibilities can be important in practice. For some applications, it may be the case that we have good theoretical reason to believe that (1) Assumption 2 is valid because those in the treatment group would, on average, do no better or no worse under the control than those in the control group, and (2) Assumption 1 is invalid because those in the control group would not do nearly as well under the treatment as those in the treatment group. Under this scenario, the naive estimator will deliver an unbiased and consistent estimate of the average treatment effect for the treated, even though it is still biased and inconsistent for both the average treatment effect for the untreated and the unconditional average treatment effect.

Now, return to the case in which neither Assumption 1 nor Assumption 2 is true. If the naive estimator is therefore biased and inconsistent for the typical average causal effect of interest, what can be done? The first recourse is to attempt to partition the sample into subgroups within which assumptions such as Assumptions 1 and/or 2 can be defended. The strategy amounts to conditioning on one or more variables that identify such strata and then asserting that the naive estimator is unbiased and consistent within these strata for one of the average treatment effects. One can then average estimates from these strata in a reasonable way to generate the average causal effect estimate of interest. We turn, in the next part of the book, to the two major conditioning strategies – matching and regression analysis – for estimating average causal effects when the naive estimator is biased and inconsistent.

2.7 Conclusions

In this chapter, we have introduced the main components of the counterfactual model of causality, also known as the potential outcome model. To motivate the presentation of matching and regression in the next part of the book, we first reintroduce causal graphs and the notational framework for modeling the treatment assignment mechanism in the next chapter. Although we will then show that matching and regression share many connections, we also aim to demonstrate that they are typically motivated in two entirely different ways, as, in the first case, an attempt to balance the variables that predict treatment assignment/selection and as, in the second case, an attempt to condition on all other relevant direct causes of the outcome. The causal graphs show this connection clearly, and hence we begin by showing how conditioning strategies represent an attempt to eliminate all net associations between the causal variable and the outcome variable that are produced by back-door paths that confound the causal effect of interest.

Appendix A: Population and Data Generation Models

In the counterfactual tradition, no single agreed-on way to define the population exists. In a recent piece, for example, Rubin (2005:323) introduces the primary elements of the potential outcome model without taking any particular position on the nature of the population, writing that "'summary' causal effects can also be defined at the level of collections of units, such as the mean unit-level causal effect for all units." As a result, a variety of possible population-based (and "collection"-based) definitions of potential outcomes, treatment assignment patterns, and observed outcomes can be used. In this appendix, we explain the choice of population model that we will use throughout the book (and implicitly, unless otherwise specified).

Because we introduce populations, samples, and convergence claims in this chapter, we have placed this appendix here. Nonetheless, because we have not yet introduced models of causal exposure, some of the fine points in the following discussion may well appear confusing (notably how "nature" performs randomized experiments behind our backs). For readers who wish to have a full understanding of the implicit superpopulation model we will adopt, we recommend a quick reading of this appendix now and then a second more careful reading after completing Chapters 3 and 4.

Our Implicit Superpopulation Model

The most expedient population and data generation model to adopt is one in which the population is regarded as a realization of an infinite superpopulation. This setup is the standard perspective in mathematical statistics, in which random variables are assumed to exist with fixed moments for an uncountable and unspecified universe of events. For example, a coin can be flipped an infinite number of times, but it is always a Bernoulli distributed random variable for which the expectation of a fair coin is equal to .5 for both heads and tails. For this example, the universe of events is infinite because the coin can be flipped forever.

Many presentations of the potential outcome framework adopt this basic setup, presumably following Rubin (1977) and Rosenbaum and Rubin (1983b, 1985a). For a binary cause, potential outcomes Y^1 and Y^0 are implicitly assumed to have expectations $E[Y^1]$ and $E[Y^0]$ in an infinite superpopulation. Individual realizations of Y^1 and Y^0 are then denoted y_i^1 and y_i^0. These realizations are usually regarded as fixed characteristics of each individual i.

This perspective is tantamount to assuming a population machine that spawns individuals forever (i.e., the analog to a coin that can be flipped forever). Each individual is born as a set of random draws from the distributions of Y^1, Y^0, and additional variables collectively denoted by S. These realized values y^1, y^0, and s are then given individual identifiers i, which then become y_i^1, y_i^0, and s_i.

The challenge of causal inference is that nature also performs randomized experiments in the superpopulation. In particular, nature randomizes a causal variable D within strata defined by the values of S and then sets the value of Y as y_i equal to y_i^1 or y_i^0, depending on the treatment state that is assigned to each individual. If nature assigns an individual to the state $D = 1$, nature then sets y_i equal to y_i^1. If nature assigns an individual to the state $D = 0$, nature then sets y_i equal to y_i^0. The differential probability of being assigned to $D = 1$ instead of $D = 0$ may be a function in S, depending on the experiment that nature has decided to conduct (see Chapters 3 and 4). Most important, nature then deceives us by throwing away y_i^1 and y_i^0 and giving us only y_i.

In our examples, a researcher typically obtains data from a random sample of size N from a population, which is in the form of a dataset $\{y_i, d_i, s_i\}_{i=1}^N$. The sample that generates these data is drawn from a finite population that is itself only one realization of a theoretical superpopulation. Based on this set-up, the joint probability distribution in the sample $\Pr_N(Y, D, S)$ must converge in probability to the true joint probability distribution in the superpopulation $\Pr(Y, D, S)$ as the sample size approaches infinity. The main task for analysis is to model the relationship between D and S that nature has generated in order use observed data on Y to estimate causal effects defined by Y^1 and Y^0.

Because of its expediency, we will usually write with this superpopulation model in the background, even though the notions of infinite superpopulations and sequences of sample sizes approaching infinity are manifestly unrealistic. We leave the population and data generation model largely in the background in the main text, so as not to distract the reader from the central goals of our book.

Alternative Perspectives

There are two main alternative models of the population that we could adopt. The first, which is consistent with the most common starting point of the survey sampling literature (e.g., Kish 1965), is one in which the finite population is recognized as such but treated as so large that it is convenient to regard it as infinite. Here, values of a sample statistic (such as a sample mean) are said to equal population values in expectation, but now the expectation is taken over repeated samples from the population (see Thompson 2002 for an up-to-date accounting of this perspective). Were we to adopt this perspective, rather than our superpopulation model, much of what we write would be the same. However, this perspective tends to restrict attention to large survey populations (such as all members of the U.S. population older than 18) and makes it cumbersome to discuss some of the estimators we will consider (e.g., in Chapter 4, where we will sometimes define causal effects only across the common support of some random variables, thereby necessitating a redefinition of the target population).

The second alternative is almost certainly much less familiar to many empirical social scientists but is a common approach within the counterfactual causality literature. It is used often when no clearly defined population exists from which the data can be said to be a random sample (such as when a collection

of data of some form is available and an analyst wishes to estimate the causal effect for those appearing in the data). In this situation, a dataset exists as a collection of individuals, and the observed individuals are assumed to have fixed potential outcomes y_i^1 and y_i^0. The fixed potential outcomes have average values for those in the study, but these average values are not typically defined with reference to a population-level expectation. Instead, analysis proceeds by comparison of the average values of y_i for those in the treatment and control groups with all other possible average values that could have emerged under all possible permutations of treatment assignment. This perspective then leads to a form of randomization inference, which has connections to exact statistical tests of null hypotheses most commonly associated with Fisher (1935). As Rosenbaum (2002) shows, many of the results we present in this book can be expressed in this framework (see also Rubin 1990, 1991). But the combinatoric apparatus required for doing so can be cumbersome (and at times requires constraints, such as homogeneity of treatment effects, that are restrictive). Nonetheless, because the randomization inference perspective has some distinct advantages in some situations, we will refer to it at several points throughout the book. And we strongly recommend that readers consult Rosenbaum (2002) if the data under consideration arise from a sample that has no straightforward and systematic connection to a well-defined population. In this case, sample average treatment effects may be the only well-defined causal effects, and, if so, then the randomization inference tradition is a clear choice.

Appendix B: Extension of the Framework to Many-Valued Treatments

In this chapter, we have focused discussion mostly on binary causal variables, conceptualized as dichotomous variables that indicate whether individuals are observed in treatment and control states. As we show here, the counterfactual framework can be used to analyze causal variables with more than two categories.

Potential and Observed Outcomes for Many-Valued Treatments

Consider the more general setup, in which we replace the two-valued causal exposure variable, D, and the two potential outcomes Y^1 and Y^0 with (1) a set of J treatment states, (2) a corresponding set of J causal exposure dummy variables, $\{Dj\}_{j=1}^J$, and (3) a corresponding set of J potential outcome random variables, $\{Y^{Dj}\}_{j=1}^J$. Each individual receives only one treatment, which we denote Dj^*. Accordingly, the observed outcome variable for individual i, y_i, is then equal to $y_i^{Dj^*}$. For the other $J-1$ treatments, the potential outcomes of individual i exist in theory as $J-1$ other potential outcomes y_i^{Dj} for $j \neq j^*$, but they are counterfactual.

Consider the fundamental problem of causal inference for many-value treatments presented in Table 2.4 (which is simply an expansion of Table 2.1 to

Table 2.4: The Fundamental Problem of Causal Inference for Many-Valued Treatments

Group	Y^{D1}	Y^{D2}	...	Y^{DJ}
Takes $D1$	Observable as Y	Counterfactual	...	Counterfactual
Takes $D2$	Counterfactual	Observable as Y	...	Counterfactual
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Takes DJ	Counterfactual	Counterfactual	...	Observable as Y

many-valued treatments). Groups exposed to alternative treatments are represented by rows with, for example, those who take treatment $D2$ in the second row. For a binary treatment, we showed earlier that the observed variable Y contains exactly half of the information contained in the underlying potential outcome random variables. In general, for a treatment with J values, Table 2.4 shows that the observed outcome variable Y contains only $1/J$ of the total amount of information contained in the underlying potential outcome random variables. Thus, the proportion of unknown and inherently unobservable information increases as the number of treatment values, J, increases.

For an experimentalist, this decline in the relative amount of information in Y is relatively unproblematic. Consider an example in which a researcher wishes to know the relative effectiveness of three pain relievers for curing headaches. The four treatments are "Take nothing," "Take aspirin," "Take ibuprofen," and "Take acetaminophen." Suppose that the researcher rules out an observational study, in part because individuals have constrained choices (i.e., pregnant women may take acetaminophen but cannot take ibuprofen; many individuals take a daily aspirin for general health reasons). Instead, she gains access to a large pool of subjects not currently taking any medication and not prevented from taking any of the three medicines.[19] She divides the pool randomly into four groups, and the drug trial is run. Assuming all individuals follow the experimental protocol, at the end of the data collection period the researcher calculates the mean length and severity of headaches for each of the four groups.

Even though three quarters of the cells in a 4×4 observability table analogous to Table 2.4 are counterfactual, she can effectively estimate the relative effectiveness of each of the drugs in comparison with each other and in comparison with the take-nothing control group. Subject to random error, contrasts such as $E_N[y_i|\text{Take aspirin}] - E_N[y_i|\text{Take ibuprofen}]$ reveal all of the average treatment effects of interest. The experimental design allows her to ignore the counterfactual cells in the observability table by assumption. In other words, she can assume that the average counterfactual value of Y^{Aspirin} for those who

[19] Note that, in selecting this group, she has adopted a definition of the population of interest that does not include those who (1) take one of these pain relievers regularly for another reason and (2) do not have a reason to refuse to take one of the pain relievers.

Table 2.5: The Observability Table for Estimating how Education Increases Earnings

Education	Y^{HS}	Y^{AA}	Y^{BA}	Y^{MA}
Obtains HS	Observable as Y	Counterfactual	Counterfactual	Counterfactual
Obtains AA	Counterfactual	Observable as Y	Counterfactual	Counterfactual
Obtains BA	Counterfactual	Counterfactual	Observable as Y	Counterfactual
Obtains MA	Counterfactual	Counterfactual	Counterfactual	Observable as Y

took nothing, took ibuprofen, and took acetaminophen (i.e., $E[Y^{\text{Aspirin}}|\text{Take nothing}]$, $E[Y^{\text{Aspirin}}|\text{Take ibuprofen}]$, and $E[Y^{\text{Aspirin}}|\text{Take acetaminophen}]$) can all be assumed to be equal to the average observable value of Y for those who take the treatment aspirin, $E[Y|\text{Take aspirin}]$. She can therefore compare sample analogs of the expectations in the cells of the diagonal of the observability table, and she does not have to build contrasts within its rows. Accordingly, for this type of example, comparing the effects of multiple treatments with each other is no more complicated than the bivariate case, except insofar as one nonetheless has more treatments to assign and resulting causal effect estimates to calculate.

Now consider a variant on the education-earnings example from the first chapter. Suppose that a researcher hopes to estimate the causal effect of different educational degrees on labor market earnings, and further that only four degrees are under consideration: a high school degree (HS), an associate's degree (AA), a bachelor's degree (BA), and a master's degree (MA). For this problem, we therefore have four dummy treatment variables corresponding to each of the treatment states: HS, AA, BA, and MA. Table 2.5 has the same structure as Table 2.4. Unlike the pain reliever example, random assignment to the four treatments is impossible. Consider the most important causal effect of interest for policy purposes, $E[Y^{\text{BA}} - Y^{\text{HS}}]$, which is the average effect of obtaining a bachelor's degree instead of a high school degree.

Suppose that an analyst has survey data on a set of middle-aged individuals for whom earnings at the most recent job and highest educational degree is recorded. To estimate this effect without asserting any further assumptions, the researcher would need to be able to consistently estimate population-level analogs to the expectations of all of the cells of Table 2.5 in columns 1 and 3, including six counterfactual cells off of the diagonal of the table. The goal would be to formulate consistent estimates of $E[Y^{\text{BA}} - Y^{\text{HS}}]$ for all four groups of differentially educated adults. To obtain a consistent estimate of $E[Y^{\text{BA}} - Y^{\text{HS}}]$, the researcher would need to be able to consistently estimate $E[Y^{\text{BA}} - Y^{\text{HS}}|HS = 1]$, $E[Y^{\text{BA}} - Y^{\text{HS}}|AA = 1]$, $E[Y^{\text{BA}} - Y^{\text{HS}}|BA = 1]$, and $E[Y^{\text{BA}} - Y^{\text{HS}}|MA = 1]$, after which these estimates would be averaged across the distribution of educational attainment. Notice that this requires the consistent estimation of some doubly counterfactual contrasts, such as the effect on earnings of shifting from

a high school degree to a bachelor's degree for those who are observed with a master's degree. The researcher might boldly assert that the wages of all high school graduates are, on average, equal to what all individuals would obtain in the labor market if they instead had high school degrees. But this is very likely to be a mistaken assumption if it is the case that those who carry on to higher levels of education would have been judged more productive workers by employers even if they had not attained more than high school degrees.

As this example shows, a many-valued treatment creates substantial additional burden on an analyst when randomization is infeasible. For any two-treatment comparison, one must find some way to estimate a corresponding $2(J-1)$ counterfactual conditional expectations, because treatment contrasts exist for individuals in the population whose observed treatments place them far from the diagonal of the observability table.

If estimating all of these counterfactual average outcomes is impossible, analysis can still proceed in a more limited fashion. One might simply define the parameter of interest very narrowly, such as the average causal effect of a bachelor's degree only for those who typically attain high school degrees: $E[Y^{\text{BA}} - Y^{\text{HS}}|\text{HS} = 1]$. In this case, the causal effect of attaining a bachelor's degree for those who typically attain degrees other than a high school degree are of no interest for the analyst.

Alternatively, there may be reasonable assumptions that one can invoke to simplify the complications of estimating all possible counterfactual averages. For this example, many theories of the relationship between education and earnings suggest that, for each individual i, $y_i^{\text{HS}} \leq y_i^{\text{AA}} \leq y_i^{\text{BA}} \leq y_i^{\text{MA}}$. In other words, earnings never decrease as one obtains a higher educational degree. Asserting this assumption (i.e., taking a theoretical position that implies it) may allow one to ignore some cells of the observability table that are furthest from the direct comparison one hopes to estimate.

Other Aspects of the Counterfactual Model for Many-Valued Treatments

Aside from the expansion of the number of causal states, and thus also treatment indicator variables and corresponding potential outcome variables, all other features of the counterfactual model remain essentially the same. SUTVA must still be maintained, and, if it is unreasonable, then more general methods must again be used to model treatment effects that may vary with patterns of treatment assignment. Modeling treatment selection remains the same, even though the added complexity of having to model movement into and out of multiple potential treatment states can be taxing. And the same sources of bias in standard estimators must be considered, only here again the complexity can be considerable when there are multiple states beneath each contrast of interest.

To avoid all of this complexity, one temptation is to assume that treatment effects are linear additive in an ordered set of treatment states. For the effect of education on earnings, a researcher might instead choose to move forward under the assumption that the effect of education on earnings is linear additive

in the years of education attained. For this example, the empirical literature has demonstrated that this is a particularly poor idea. For the years in which educational degrees are typically conferred, individuals appear to receive an extra boost in earnings. When later discussing the estimation of treatment effects using linear regression for many-valued treatments, we will discuss a piece by Angrist and Krueger (1999) that shows very clearly how far off the mark these methods can be when motivated by unreasonable linearity and additivity assumptions.