Chapter 2

Counterfactuals and the Potential Outcome Model

In this chapter, we introduce the foundational components of the potential outcome model. We first discuss causal states, the relationship between potential and observed outcome variables, and the usage of the label "counterfactual" to refer to unobserved potential outcomes. We introduce average causal effects and then discuss the assumption of causal effect stability, which is maintained explicitly in most applications that use the potential outcome model. We discuss simple estimation techniques and demonstrate the importance of considering the relationship between the potential outcomes and the process of causal exposure. We conclude by extending our presentation to over-time potential outcome variables for one or more units of analysis, as well as causal variables that take on more than two values.

2.1 Defining the Causal States

The counterfactual framework for observational data analysis presupposes the existence of well-defined causal states to which all members of the population of interest could be exposed.[1] As we will show in the next section, causal effects are then defined based on comparisons of outcomes that would result from exposure to alternative causal states. For a binary cause, the 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.

Although these labels are simple, the assumed underlying states must be very carefully defined so that the contribution of an empirical analysis based upon them is clear. Some of the relevant issues can only be discussed as we introduce additional

[1] We justify the importance of carefully defining the boundaries of the population of interest when presenting average causal effects later in this chapter. 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.

pieces of the full counterfactual framework in this chapter and the next – moving from the definition of individual-level causal effects, through average causal effects, and then to causal graphs and the underlying structural equations that they represent. Nonetheless, some initial clarification of core definitional issues for these causal states is essential.

Fine Articulation. To appreciate the value of finely articulated causal states, consider the examples introduced in Section 1.3. The worker training example is straightforward, and the two states are "entered a training program" (the treatment state) and "did not enter a training program" (the control state). The charter school example is similar. Here, the alternative states are "enrolled in a charter school" (the treatment state) and "enrolled in a regular public school" (the control state, although possibly referred to as an alternative treatment state). One possible complication with these examples is the possibility of inherent differences across training programs, charter schools, and regular public schools. If any such treatment-site heterogeneity exists, then stratified analyses may be necessary, perhaps by regions of the country, size of the program or school, or whatever other dimension suggests that variability of the causal states deserves explicit modeling.[2]

Other examples, at least as executed in the extant research, have causal states that are not finely articulated. Consider the classic political participation line of inquiry. For the relationship between socioeconomic status and political participation, there are many underlying 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 internally differentiated concept of socioeconomic status, which social scientists have created for their own analytic purposes. It is therefore unsurprising that some of the most recent literature (e.g., Henderson and Chatfield 2011; Kam and Palmer 2008) has specified more finely articulated causal states for this line of research, such as "entered college" (the treatment state) in comparison to "did not enter college" (the control state).

The father absence example is an intermediate case. The original research attempted to estimate the broad effects of single parenthood on the outcomes of children and adolescents (see McLanahan and Sandefur 1994; Wu and Wolfe 2001). The more recent literature has focused on the narrowly defined treatment state of father absence. Even so, much variation remains in both the definition of this treatment state and the relevant comparison (or control) state, as noted in a review piece:

> Studies in this field measured father absence in several ways, which the reader should keep in mind when interpreting and comparing results across studies. Some studies compared children of divorced parents with children of stably married parents; others compared children whose parents married after their childs birth with those parents who never married.... More recently, researchers have started to use even more nuanced categories to

[2]For example, Hong and Raudenbush (2006) provide a careful analysis of retention policies in primary education, implementing this type of treatment-site stratification based on the average level of retention in different public schools. We will discuss these types of studies in Section 2.5.

> measure family structure – including married biological-parent families, cohabiting biological-parent families, married stepparent families, cohabiting stepparent families, and single parents by divorce and nonmarital birth – reflecting the growing diversity of family forms in society.... We did not identify any studies that used causal methods to study the effects of same-sex unions. (McLanahan et al. 2013:408)

In general, research that takes account of heterogeneity by splitting treatment states into mutually exclusive component states will break new ground if sufficient data are available to estimate the more narrowly defined treatment effects.

Nominal States from Constitutive Features. We take a pragmatic but principled position on the characteristics of causal states, and in this subsection we want to clarify our position for readers who are interested in debates on the nature of causation in philosophy and how those debates are relevant for social science research. Readers who are uninterested in these debates may wish to skim this subsection now and reengage it after reading Chapter 10 on causal mechanisms and Section 13.1 on objections to the counterfactual approach to observational data analysis (pages 438–446). In fact, most scholars who work with counterfactual models in social science research do not take any positions on the issues that we raise in this subsection, and their research shows that much useful work can proceed by taking and using the causal states as measured, without considering the features that give them their capacities to generate effects.

Having offered these warnings, we will now explain why we take the position that each state of each treatment should be regarded as a nominal state with constitutive features (e.g., entities, activities, and relations) that are jointly capable of producing the outcome of interest. Consider the Catholic school example, where the nominal states are "enrolled in a Catholic school" and "enrolled in a public school" and where the outcome is "learning." Each type of school has teachers, classrooms, curricula, administrators, normative environments, affiliated institutions, and networks of peers and parents. The literature on the differences between Catholic schools and public schools suggests that these constitutive features of the schools are interrelated in ways that differ by type of school. Accordingly, while Catholic schools and public schools both produce student learning, the particular ways in which they do so are thought to differ meaningfully across type of school, and in ways that have not been documented comprehensively with available data. Nonetheless, we can still conceive of each student in the population of interest being exposed to each type of school, and we can assume that each student would then experience the learning generated in toto by the joint capacities of the constituent features of each type of school.

Taking this position, while at the same time embracing counterfactual dependence, implies that we see value in mounting causal analysis in the social sciences on top of a foundation that conjoins a metaphysics of causal powers with a metaphysics of counterfactual dependence (see Collins, Hall, and Paul 2004; Mumford and Anjum 2011). The price for such an inclusive pragmatism is an elaborate metaphysics, which most philosophers would likely regard as insufficiently elegant and insufficiently reductive. With reference to Hume's example of billiard balls (Hume 1977[1772]), our position requires that we adopt the following specific (but perhaps painfully elaborate) account

of the nature of causation: The cue ball causes the second billiard ball to roll a particular observed distance because billiard balls are spheres *and* because the cue ball was struck by the player's pool cue in such a way that it then struck the second billiard ball at a particular angle and with a particular force. Furthermore, the cue ball would not have caused the second billiard ball to roll the same observed distance if the billiard balls had instead not been spheres or if the cue ball had not been struck by the player's pool cue in the exact same way. Thus, the causal effect of the cue ball on the second billiard ball is a joint product of the spherical feature of the billiard balls as well as the external intervention of the pool player.[3]

For the sorts of social science examples we consider in this book, we will express the effects of causal states using contrasts between observed exposure to one state and what-if counterfactual exposure to another state. However, we will also take the position that any such claims about the effects of exposure to alternative states are incomplete until those claims are accompanied by accounts of the constitutive features of the causal states and how those features are thought to grant the states the power to generate outcomes.[4] The most complete accounts point to evidence that mechanisms exist that are capable of generating the outcomes of interest (and, better yet, that it is reasonable to believe that these mechanisms will be able to explain why exposure to alternative causal states generates differences).

Consider an example where many of these issues are settled. For the Catholic school effect, analysis can proceed within a guiding framework shaped by a rich background literature. The historical events that generated public schools and Catholic schools as coherent institutions suggests that they can be meaningfully compared when studying student achievement because they each aim to produce learning for core academic subjects, even though they each pursue additional distinct goals (see Tyack 1974 for one of the most widely read accounts). In addition, each type of school has a rich set of literature that has examined the mechanisms that generate learning. For Catholic

[3]With the goal of reducing the complexity of such an account, philosophers seem inclined to take positions on whether the spherical characteristic of the cue ball (its "causal power") is more fundamental than the striking (i.e., the "counterfactual dependence" induced by the intervention of the player), whether "striking" can be defined in the absence of an intervening pool player, whether a valid explanation can simply be deduced from laws of motion in space and time, whether anything is transferred between the two billiard balls at the moment of impact, and so on. We see no reason to take a position on these matters in this book, and we therefore quite consciously violate rule 4 of Paul and Hall (2013:40), "Thou shall not be an ontological wimp."

[4]We see little value in placing restrictions on what types of origination accounts are admissible and should be relied upon. For some causal claims, historical narratives are appropriate, to the extent that they focus on especially salient institutional histories while pushing into the background the multitude of specific decisions of all individuals that have given shape to the constitutive features of the alternative states (see Reed 2011 for examples of such "forming" narratives). In other cases, the origins of the states can be explained as contrasting values for built concepts, based on underlying analytic dimensions drawn from the extant social science literature, where these underlying dimensions have been chosen precisely because background evidence exists that they are sources of productive causal power of the nominal causal states of interest (see Goertz 2006 for examples). However, we see one complication for this second type of account, as foreshadowed by our discussion of socioeconomic status above. Causal states drawn from values for a built concept may be real only in the minds of researchers. As a result, explanations based upon them may appear nonsensical to individuals who are purported to be producing the effects of interest. Whether such behind-the-back accounts are to be regarded as powerful or not is almost certainly a domain-specific consideration, which will also vary with the goals of a study.

schools, several complementary narratives exist that provide arguments that suggest why Catholic schools have the capacity to be more effective than public schools. These narratives include those that emphasize the relations embedded in parental network structure alongside an appropriated ideology of the Catholic church (see Coleman and Hoffer 1987), those that emphasize the extent to which Catholic schools are especially responsive to parental feedback and the threat of exit (Chubb and Moe 1990), and those that emphasize the trusting relationships between teachers and administrators that flow from a shared purpose (Bryk, Lee, and Holland 1993).

Of course, the existence of such lower-level claims about the specific mechanistic capacities of features of Catholic schools begs the question: When is it advisable to decompose nominal causal states into component causal states with their own capacities for producing the outcome of interest? We see the answer to this question as subject- and domain-specific. Accordingly, we see little value in making general arguments about when such decomposition is feasible because of the inherent separability of the productive capacities attached to particular constitutive features or is infeasible because of the deeply entangled complementarities among them. And, as we will argue later in Chapter 10, it is generally impossible to take a position on these issues in any given study without first stipulating the causal structure of the mechanisms that are presumed to produce the outcome. Fortunately, as we will demonstrate in the intervening chapters, causal effects defined only by nominal causal states can be sufficiently precise so that their estimation is itself feasible and very much worthwhile.

Local and Reasonable. Consider the literature on socioeconomic status as a fundamental cause of health and mortality, which takes as its defining feature the argument that it is only occasionally useful to identify causal states for the measurable dimensions underneath the fundamental cause of socioeconomic status (see page 19). For these scholars, it is the abundant causal pathways that link socioeconomic status with health and mortality that are most noteworthy because of the robust, total associations that they generate. Isolating a particular causal effect that is attributable to a contrast defined by two clearly defined underlying causal states embedded within socioeconomic status could still be useful, such as for the estimation of a health disparity attributable to a family income difference of $25,000. The claim of this literature is that this narrow exercise could become counterproductive if it detracted from the broader claim of fundamental causality, as would be the case if the analyst were to imply that any such narrow effect is as robust as the total causal effect that socioeconomic status exerts on health and mortality.

A fundamental-cause orientation may be useful in challenging the status quo in research areas that have become too narrowly focused on only a few relevant causal pathways, but widespread adoption of the fundamental-cause orientation to causal analysis would not be productive. In many areas of research, it would not be hard to take collections of narrowly and carefully defined causal contrasts, lump them together into latent constructs, and then assert that, over sufficiently long intervals, the latent construct is a fundamental cause because the mechanisms that are activated by component causes switch on and off over time. Indeed, considering our other examples in Section 1.3, one could argue quite easily that the socioeconomic status of one's parents is a fundamental cause of educational attainment, subsequent labor market earnings, political participation, and fertility decisions. We doubt many scholars would disagree

with such broad claims, and most would likely interpret them as consistent with conclusions drawn by scholars working with comparatively coarse data more than six decades ago. More importantly, we think it unlikely that the reassertion of such broad claims would encourage researchers to move in productive new directions. Instead, we see the counterfactual perspective, and the potential outcome model in particular, as enabling the pursuit of a more ambitious goal: the careful delineation of the relevant causal states that lie within any purported fundamental causes and then the estimation of the specific effects generated by contrasts between them. Should there be reason to expect that any such effects vary in time, then their estimation across time demands empirical analysis, not simply the assertion that such effects, by nature of their variability in time, can only be regarded as specialized instantiations of more fundamental causes.

For a related reasonableness concern, consider a specific political participation example. To what extent do restrictions on who can vote determine who wins elections? A 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?[5] 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.

A relevant criterion, although necessarily subjective, is whether it "stretches the mind" too much to imagine conceivable alternative worlds in which all else remains the same, except for the instantiation of the alternative causal states. For this particular example, the "too much" criterion was not likely crossed. Scholars in political sociology and criminology supported publication through blind peer review in the discipline's highest prestige journal, the *American Sociological Review*. Reviewers presumably saw this particular line of research as an important contribution to our knowledge on how changing laws to allow felons and ex-felons to vote could have potential effects on

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

election outcomes, and they must have concluded that there was value in understanding such effects in hypothetical isolation from other changes that would also likely co-occur in the real world along with the contemplated legislative changes.

The more general point, however, is that it is important that the "what would have been" nature of the conditionals that define the causal states of interest be carefully considered. When a facile ceteris paribus assumption is invoked to relieve the analyst from having to discuss other contrasts that are nearly certain to occur at the same time, the posited causal states may be open to the charge that they are too improbable or ill-defined to justify the pursuit of a causal analysis based on them.[6]

2.2 Potential Outcomes and Individual-Level Treatment Effects

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.

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.[7] The individual-level causal effect of the treatment

[6] The philosopher Nancy Cartwright (2007a, 2007b) would refer to an analysis that defines potential outcomes (see next section) in terms of ill-conceived causal states as generating "impostor counterfactuals." She stresses the need for full causal models of all interrelated causes of outcomes, so that the effects of causes are not too narrowly assessed. She writes:

> To evaluate counterfactuals ... we need a causal model; and the causal model must contain all the information relevant to the consequent about all the changes presumed in the antecedent. There is no other reasonable method on offer to assess counterfactuals. We may not always produce a model explicitly, but for any grounded evaluation there must be a causal model implicit; and our degree of certainty about our counterfactual judgments can be no higher than our degree of certainty that our causal model is correct. (Cartwright 2007a:193)

We agree with the value of having a causal model, as will become clear in subsequent chapters. However, Cartwright takes this position to an extreme that is counterproductive for practice; see Pearl (2009:362–65).

[7] 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,

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

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

is then defined as

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

Before proceeding, two caveats on this definition of individual-level causal effects should be noted. First, the individual-level causal effect can be defined in ways other than as the linear difference between the two relevant potential outcomes.[8] One obvious possibility is the ratio of one individual-level potential outcome to another, y_i^1/y_i^0. In some research areas, alternative definitions at the individual level may have advantages. The most prominent case is epidemiology, where the goal of estimating risk factors for health outcomes continues to dominate practice and leads to a frequent preference for ratio-based rather than difference-based comparisons. Nonetheless, the overwhelming majority of the literature represents individual-level causal effects as linear differences, as in Equation (2.1).

Second, 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 the inherent uncertainty of their beliefs, which may properly reflect true variability in their individual-level potential outcomes. But, with data for which a potential outcome is necessarily observed for any individual as a scalar value (via an observed outcome variable, defined later), 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.

2.3 Treatment Groups and Observed Outcomes

For a binary cause with two causal states and associated potential outcome variables Y^1 and Y^0, a corresponding causal exposure variable, D, is specified that 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

does seem to have proven itself in our own classes, offering the right balance between specificity and compactness.

[8] Rubin (2005, figure 1) uses the general notation "v." for "versus" to depict individual-level effects in their most general form.

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 – and not use the words treatment and control at all. Indeed, we recognize that for some readers the usage of treatment and control language may feel sufficiently heterodox relative to the semantics of the areas in which they work that avoidance of these terms seems prudent. If so, it is perfectly acceptable to adopt parallel language without using the words treatment and control.

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

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

This paired definition is often written compactly as

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

Equation (2.2) implies that 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 potential outcome model of counterfactual causality. Table 2.1 depicts the "problem," which one might alternatively refer to as the "fundamental reality of causal analysis." Causal effects are defined by contrasts within rows, which refer to groups of individuals observed 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.[9]

[9] 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

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

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.

2.4 The Average Treatment Effect

Because it is typically impossible to calculate individual-level causal effects, we usually focus attention on the estimation of carefully defined aggregate causal effects. When we adopt the linear difference in potential outcomes as the definition of the individual-level causal effect, we typically define aggregate causal effects as averages of these individual-level effects. These average causal effects can de defined for any subset of the population, and throughout this book we will consider many different average effects. In this section, we introduce the broadest possible average effect, which is the average treatment effect (ATE) in the population as a whole.

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.[10]

For Equation (2.3), the expectation is defined with reference to the population of interest. For the fertility pattern example introduced in Section 1.3, the population would be one or more birth cohorts of women in a particular country. For the election outcome examples, the population would be "all eligible voters" or "all eligible voters in Florida." For other examples, such as the worker training example, the population

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.

[10] However, at a deeper level, 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.

would be "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.[11]

Note also that the subscripting on i for the individual-level causal effect, δ_i, has been dropped for Equation (2.3). Even so, the definition of the ATE should not be interpreted to suggest that we now must assume that the treatment effect is constant in the population in any fundamental sense. Rather, we can drop the subscript i in Equation (2.3) 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.[12]

To see all of these pieces put together, consider the Catholic school example. 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. The ATE, $E[\delta]$, is then the average value among all students in the population of these what-if differences in test scores. The ATE is also equal to the expected value of the what-if difference in test scores for a randomly selected student from the population.

An alternative group-level causal effect that we will not consider much in this book is the causal risk ratio,

$$\frac{Pr[Y^1 = 1]}{Pr[Y^0 = 1]}, \tag{2.4}$$

where now the outcomes Y^1 and Y^0 are indicator variables equal to 1 if the outcome of interest is present and 0 if not. This group-level effect is the analog to the individual-level ratio of potential outcomes, y_i^1/y_i^0, noted earlier in Section 2.2. The causal risk ratio is most frequently analyzed in epidemiology and the health sciences, where risk-factor analysis remains dominant and the outcomes are typically onset of a disease or a troubling symptom thereof (see Hernán and Robins 2006a). For our purposes, most outcomes modeled as causal risk ratios can be translated to average treatment effects, interpreting $E[Y^1] - E[Y^0]$ as $Pr(Y^1 = 1) - Pr(Y^0 = 1)$. Expectations of indicator variables are equivalent to probabilities of indicator variables, and an interval metric

[11] 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.

[12] For example, at many times in the book, we will stress that quantities such as the ATE should not be assumed to be equal to the individual-level causal effect for any individual i, which we will express as $\delta_i \neq E[\delta_i] = E[\delta]$ for all i. In words, when individual-level heterogeneity of causal effects is present, individual-level causal effects, δ_i, will not all be equal to the average of these individual-level causal effects, $E[\delta_i]$, which is, by the definition of the expectation operator, equal to $E[\delta]$.

is at least as sensible as a ratio metric for all of the examples we will consider. (The ratio metric might be preferable if we were attempting to make effect comparisons across outcomes with very different base rates, such as the effect of the same treatment on pancreatic cancer and hypertension.)

2.5 The Stable Unit Treatment Value Assumption

In most applications, the potential outcome model retains its tractability through the maintenance of a strong assumption known as the stable unit treatment value assumption or SUTVA (see Rubin 1980b, 1986). In economics, a version of this assumption is sometimes referred to as a no-macro-effect or partial equilibrium assumption (see Garfinkel, Manski, and Michalopoulos 1992, Heckman 2000, 2005, for the history of these ideas, and Manski and Garfinkel 1992 for examples).[13]

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 changes in the treatment exposures of all 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.[14] 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.

[13] SUTVA is a much maligned acronym, and many others use different labels. Manski (2013a:S1), for example, has recently labeled the same assumption the "individualistic treatment response" assumption in order "to mark it as an assumption that restricts the form of treatment response functions."

[14] For this example, assume that the values of y_i^1 and y_i^0 for each individual i are either deterministic potential outcomes or exactly equal to $E[Y_i^1]$ and $E[Y_i^0]$ for each individual i. Also, assume that these three subjects comprise a perfectly representative sample of the population.

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

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 of each individual as functions across all potential configurations of the elements of vector $\mathbf{d}$. Accordingly, the outcome for individual i under the treatment is $y_i^1(\mathbf{d})$, and the outcome for individual i under the control is $y_i^0(\mathbf{d})$. The treatment effect for each individual i is then

$$\delta_i(\mathbf{d}) = y_i^1(\mathbf{d}) - y_i^0(\mathbf{d}). \tag{2.5}$$

With this more general setup, individual-level treatment effects could be different for every possible pattern of treatment exposure.

SUTVA is what allows us to declare $y_i^1(\mathbf{d}) = y_i^1$ and $y_i^0(\mathbf{d}) = y_i^0$ and, as a result, assert that individual-level causal effects δ_i exist that are independent of the overall configuration of causal exposure. If SUTVA cannot be maintained, then the simplified definition in Equation (2.1) is invalid, and the individual-level treatment effect must be written in its most general form in Equation (2.5), with all ensuing analysis proceeding conditional on alternative vectors $\mathbf{d}$.

Sometimes it is argued that SUTVA is so restrictive that we need an alternative conception of causality for the social sciences. Our position is that SUTVA reveals the limitations of social science data and the perils of immodest causal modeling rather than the limitations of the potential outcome 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 all causal effect estimates offered. Such reflection may lead one to determine that only the more general case of the potential outcome framework can be justified, and this may necessitate building the analysis on top of the individual-level treatment effect defined in Equation (2.5) rather than the SUTVA-simplified variant in Equation (2.1). In some cases, however, analysis can proceed assuming SUTVA, as long as all resulting estimates are given restricted interpretations, as we now explain.

Typical SUTVA violations share two interrelated features: (1) influence patterns that result from contact across individuals in social or physical space and (2) dilution/concentration patterns that one can assume would result from changes in the prevalence of the treatment. Neither feature is entirely distinct from the other, and in many cases dilution/concentration effects arise because influence patterns are present. Yet, if the violation can be interpreted as a dilution/concentration pattern, even when generated in part by an underlying influence pattern, then the analyst can proceed by scaling back the asserted relevance of any estimates to situations where the prevalence of the treatment is not substantially different.

For a simple example, consider the worker training example. Here, the plausibility 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.

For an example of where influence patterns are more of a threat to SUTVA, consider the example of the Catholic school effect. 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 the pattern of effects would change 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 remains 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. As a result, the implications of research on the Catholic school effect for research on school voucher programs may be quite limited, and this has not been clearly enough recognized by some (see Howell and Peterson 2002, chapter 6). A similar argument applies to research on charter school effects.

Consider a SUTVA violation for a related example: 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.[15] 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 conditional 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 comparatively small).

In general, if SUTVA is maintained but there is some doubt about its validity because dilution or concentration patterns would emerge under shifts in treatment prevalence, then certain types of marginal effect estimates can usually still be defended. The idea here is to state that the estimates of average causal effects hold only for what-if movements of relatively small numbers of individuals from one hypothetical treatment state to another.

If, however, influence patterns are inherent to the causal process of interest, and the SUTVA violation cannot be considered as a type of dilution or concentration, then it will generally not be possible to circumvent the SUTVA violation by proceeding with the same analysis and only offering cautious and conditional interpretations. The most well-developed literature on situations such as these is the literature on the effects of vaccine programs (see Hudgens and Halloran 2008). Here, additional causal effects of interest using the potential outcome framework have been defined, conditional on the overall pattern of treatment assignment:

> The *indirect effect* of a vaccination program or strategy on an individual is the difference between what the outcome is in the individual not being vaccinated in a community with the vaccination program and what the outcome would have been in the individual, again not being vaccinated, but in a comparable community with no vaccination program. It is, then, the effect of the vaccination program on an individual who was not vaccinated. The combined *total effect* in an individual of being vaccinated and the vaccination program in the community is the difference between the outcome in the individual being vaccinated in a community with the vaccination program and what the outcome would be if the individual were not vaccinated and the community did not have the vaccination program. The total effect, then, is the effect of the vaccination program combined with the effect of the person having been vaccinated. The *overall effect* of a vaccination program is the difference in the outcome in an average individual

[15] 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."

> in a community with the vaccination program compared to an average individual in a comparable population with no vaccination program. (Halloran, Longini, and Struchiner 2010:272; italics in the original)

Effectively estimating these types of effects generally requires a nested randomization structure, wherein (1) vaccine programs are randomly assigned to a subset of participating groups and then (2) vaccinations are randomly given to individuals within groups enrolled in vaccine programs. These particular study designs are not possible for most social science applications, but the basic interpretive framework has been adopted to clarify what can be learned from social experiments, in particular, the Moving to Opportunity neighborhood experiment (see Sobel 2006).[16]

Much observational research on social influence patterns proceeds without consideration of these sorts of complications. Consider the contentious literature on whether peer effects have accelerated the obesity epidemic, as presented in Section 1.3 (see page 26). As we noted there, the basic claim of Christakis and Fowler (2007) is that having a friend who becomes obese increases one's own odds of becoming obese. Yet, their full set of claims is substantially more detailed, suggesting that these peer effects travel across network paths of length three before dying out. In particular, one's odds of becoming obese also increase if friends of friends become obese and if friends of friends of friends become obese. The sizes of these three effects diminish with the length of friendship distance.

Now consider whether SUTVA is reasonable for such a schedule of effects across network ties. Holding the social network structure fixed, if obesity increases in the population, then, on average, individuals have more obese friends, more obese friends of friends, and more obese friends of friends of friends. Most theoretical predictions would suggest that the effects on one's own odds of becoming obese that result from having friends of friends of friends who become obese should decline with the proportion of one's own friends who are already obese or who have just become obese.[17] Effects that cascade in these conditional ways, because they are defined across a pattern of interpersonal contact between units, nearly always violate SUTVA.[18]

[16] Suitable models for observational data are an active frontier of research (see Hong and Raudenbush 2013; Manski 2013a). Tchetgen Tchetgen and VanderWeele (2010) show that some estimators may be effective for applications with observational data if all relevant patterns of treatment assignment (i.e., $\mathbf{d}$) can be attributed to measured treatment-level variables.

[17] This means that, even if the issues raised by critics on the severity of homophily bias are invalid (see VanderWeele 2011b for a convincing case that they have been exaggerated), the pattern of effects only holds under the prevalence of obesity in the data analyzed, which is the pattern of obesity in Framingham, Massachusetts, among adults born in 1948 for whom data was collected between 1971 and 1999 (and for a social network structure elicited by an unconventional name generator). The overall pattern of declining effects may be valid, but the relation of the various lagged regression coefficients offered to well-defined causal effects of general interest may be rather thin.

[18] When we have conveyed this point to network analysis researchers, a common reaction is that the potential outcome model must not, therefore, be suitable for studying causal effects that propagate across networks. The logic of this position eludes us for two reasons. First, the potential outcome model cannot be deemed inappropriate because it makes clear how hard it is to define and estimate the effects that analysts claim that they wish to estimate. Second, the potential outcome model can accommodate SUTVA violations, although not without considerable additional effort. Weihua An (2013) demonstrates the value of counterfactual thinking for modeling peer effects, fully embedded within a social network perspective (see also VanderWeele and An 2013).

2.6 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 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.[19] This independence assumption is usually written as

$$(Y^0, Y^1) \perp\!\!\!\perp D, \tag{2.6}$$

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

This way of thinking about randomized experiments and potential outcomes can be confusing to social scientists who work primarily with observational data. The independence relationships represented by Equation (2.6) 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, because Equation (2.6) does not imply that D is independent of Y. Equation (2.6) implies only that in the full population, ex ante to any pattern of treatment assignment, D is independent of Y^0, Y^1, and any causal effects defined from them. Only after a study is undertaken

[19] As we will discuss in detail in later chapters, the word "ignorability" has a very specific meaning that is broader than implied in this paragraph. In short, ignorability also holds in the weaker situation in which S is a set of observed variables that 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.

do values for Y emerge, from $Y = DY^1 + (1 - D)Y^0$ in Equation (2.2). If individuals are randomly assigned to both the treatment and the control states, and individual causal effects are nonzero, then Y and D will be dependent because the average value of DY^1 will not be equal to the average value of $(1 - D)Y^0$.

Now consider the additional challenges posed by observational data analysis. These challenges to causal inference 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.[20] (Italics in the original)

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 (2.6). 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 are selected into 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 the expectations that determine their choices). 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.

2.7 Average Causal Effects and Naive Estimation

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

[20] Note that Rosenbaum's definition is consistent with the Cox and Reid definition quoted in Chapter 1 (see page 7).

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.

2.7.1 Conditional Average Treatment Effects

The unconditional average treatment effect, which is typically labeled the ATE in the counterfactual tradition, was defined in Equation (2.3) as $E[\delta] = E[Y^1 - Y^0]$. This average 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 broad foundational examples introduced in Section 1.3.1, such as 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

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1], \end{aligned} \tag{2.7}$$

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

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0], \end{aligned} \tag{2.8}$$

where, as for the ATE 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 ATT 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 ATE and the ATT 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 ATT in Equation (2.7) 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 ATT 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 ATC 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 average treatment effect might be of interest: $E[\delta|D = 0,$ CurrentSchool = Struggling], where of course the definition of being currently educated in a struggling school would have to be clearly specified.

The worker training example is similar, in that the subject of first investigation is surely the ATT (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. 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 ATC, and hence the all-encompassing ATE, 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), (2.7), and (2.8), 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).

2.7.2 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.[21] The naive estimator is then defined as

$$\hat{\delta}_{\text{NAIVE}} \equiv E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0], \tag{2.9}$$

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 consistent or unbiased estimates of the ATE because it converges to a contrast, $E[Y|D = 1] - E[Y|D = 0]$, that is not equivalent to (and usually not equal to) any of the average causal effects defined above. To see why, first decompose the ATE in Equation (2.3) as

$$\begin{aligned} E[\delta] = &\{\pi E[Y^1|D = 1] + (1 - \pi)E[Y^1|D = 0]\} \\ &- \{\pi E[Y^0|D = 1] + (1 - \pi)E[Y^0|D = 0]\}. \end{aligned} \tag{2.10}$$

Equation (2.10) reveals that the ATE is 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.

With observational data from a random sample of the population and without introducing additional assumptions, we can compute estimates that are consistent and unbiased for only three of the five unknowns on the right-hand side of Equation (2.10). Consider π first, which we have defined as equal to $E[D]$, and which is the fixed proportion of the population that would be assigned to (or would select into) the treatment. The sample-mean estimator, $E_N[d_i]$, is consistent for π, which we write as

$$E_N[d_i] \xrightarrow{p} \pi. \tag{2.11}$$

Equation (2.11) represents the claim that, as the sample size N increases to infinity, the sample mean of the values for d_i converges to the true value of π, which we assume is a fixed population parameter equal exactly to $E[D]$.[22] 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. (Estimators with this property are defined as

[21] 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 because it allows for greater clarity in aligning sample-level and population-level conditional expectations for subsequent expressions.

[22] Again, see our appendix to this chapter on our assumed superpopulation model. We are implicitly assuming that these sequences are well defined because conditions are such that the law of large numbers is applicable.

"consistent" in the statistical literature on estimation. We can also state that $E_N[d_i]$ is unbiased for π because the expected value of $E_N[d_i]$ over repeated samples of size N from the same population is equal to π as well. However, in this book we focus primarily on the consistency of estimators.)[23]

We can offer similar claims for consistent estimators of two other unknowns in Equation (2.10):

$$E_N[y_i|d_i = 1] \xrightarrow{p} E[Y^1|D = 1], \tag{2.12}$$

$$E_N[y_i|d_i = 0] \xrightarrow{p} E[Y^0|D = 0], \tag{2.13}$$

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 compute consistent or unbiased estimates of the two remaining unknowns in Equation (2.10): $E[Y^1|D = 0]$ 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 educated in Catholic schools and the average achievement of Catholic school students if they had instead been educated in public schools.

2.7.3 The Typical Inconsistency and Bias of the Naive Estimator

In the last section, we concluded 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 a contrast, $E[Y^1|D = 1] - E[Y^0|D = 0]$, that does not necessarily equal the ATE. In this section, we show why this contrast can be uninformative about the causal effect of interest in an observational study by analyzing the typical inconsistency and bias in the naive estimator as an estimator of the ATE.[24] Consider the following rearrangement of the decomposition in

[23] Nonetheless, we will often label estimators as "consistent and unbiased" when this is true, even though we will not state the case for unbiasedness. On the one hand, estimators of fixed, finite values in the population that are consistent are necessarily also asymptotically unbiased. On the other hand, some consistent estimators are not unbiased in finite samples (most prominently, for this book, the instrumental variable estimators that we will present in Chapter 9). As with the statistical literature on point estimation, we typically interpret unbiasedness as a desirable property among the consistent estimators that we will present. If an estimator is not consistent (i.e., is inconsistent), then in practice there is little reason to further consider it (as unbiased, etc.), especially when invoking a superpopulation perspective and assuming that a dataset for a large random sample is available.

[24] An important point of this literature is that the inconsistency and bias of an estimator is a function of the target parameter that has been selected for analysis. Because there are many causal effects that can be estimated, general statements about the inconsistency and bias of particular estimators are always conditional on a clear indication of the causal effect of interest.

Equation (2.10):

$$E[Y^1|D=1]-E[Y^0|D=0]=E[\delta]+\{E[Y^0|D=1]-E[Y^0|D=0]\} \quad (2.14)$$
$$+\ (1-\pi)\{E[\delta|D=1]-E[\delta|D=0]\}.$$

The naive estimator converges to the difference on the left-hand side of this equation, and the right-hand side shows that this difference is equal to the true ATE, $E[\delta]$, plus the expectations of two potential sources of inconsistency and bias in the naive estimator.[25] The first source, $\{E[Y^0|D=1]-E[Y^0|D=0]\}$, is a *baseline bias* equal to the difference in the expected outcome in the absence of the treatment between those in the treatment group and those in the control group. The second source, $(1-\pi)\{E[\delta|D=1]-E[\delta|D=0]\}$, is a *differential treatment effect bias* equal to the expected difference in the treatment effect between those in the treatment group 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, 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 ATE, represented by $E[\delta]$ in Equations (2.3) and (2.14). Second, individuals who attend college might have been been smarter in the first place. This source of inconsistency and 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 inconsistency and 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 bachelor's degree. Suppose further that 30 percent of the population obtains a bachelor's degree, 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 bachelor's degrees would have done better in the labor market than those without bachelor's degrees in the counterfactual state in which they did not in fact obtain bachelor's degrees (i.e., on average they would have received 6 instead of 5). Likewise, those who do not obtain bachelor's

[25] 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 (2.10) 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 (2.14).

Table 2.3 An Example of Inconsistency and Bias of the Naive Estimator When the ATE Is the Causal Effect of Interest

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

degrees would not have done as well as those who did obtain bachelor's degrees in the counterfactual state in which they did in fact obtain bachelor's degrees (i.e., on average they would have received 8 rather than 10). Accordingly, the ATT is 4, whereas the ATC is only 3.[26] Finally, if the proportion of the population that has a bachelor's degree is .3, then the ATE is 3.3, which is equal to $.3(10-6)+(1-.3)(8-5)$.

Consider now the inconsistency and bias of the naive estimator. For this example, the naive estimator, as defined in Equation (2.9), would be equal to 5 for an infinite sample (or equal to 5, on average, across repeated samples). Thus, the naive estimator is inconsistent and upwardly biased for the ATE (i.e., yielding 5 rather than 3.3), the ATT (i.e., yielding 5 rather than 4), and the ATC (i.e., yielding 5 rather than 3). Equation (2.14) gives the components of the total expected bias of 1.7 for the naive estimator as an estimate of the ATE. 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 $(1-.3)(4-3)=.7$.[27]

2.7.4 Estimating Causal Effects Under Maintained Assumptions About Potential Outcomes

What assumptions suffice to enable consistent and unbiased estimation of the ATE 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 a consistent and unbiased estimate of the ATE when randomization of the treatment is infeasible.

[26] For the causal effect of education on earnings, there is debate in the recent literature on whether the ATT is larger than the ATC. Cunha and Heckman (2007) and Carneiro, Heckman, and Vytlacil (2011) offer results in support of this pattern, but Brand and Xie (2010) offer results in opposition to it.

[27] In general, the size of this expected differential treatment effect bias declines as more of the population is characterized by the ATT than by the ATC (i.e., as π approaches 1).

Consider the following two assumptions:

$$\text{Assumption 1:} \quad E[Y^1|D=1] = E[Y^1|D=0], \tag{2.15}$$

$$\text{Assumption 2:} \quad E[Y^0|D=1] = E[Y^0|D=0]. \tag{2.16}$$

If one asserts these two equalities and then substitutes into Equation (2.10), the number of unknowns is reduced from the original five parameters to the three parameters that we know from Equations (2.11)–(2.13) can be consistently estimated with data generated from a random sample of the population. If both Assumptions 1 and 2 are maintained, then the ATE, ATT, and ATC in Equations (2.3), (2.7), and (2.8), respectively, are all equal. And the naive estimator is consistent and unbiased for all of them.

When would Assumptions 1 and 2 in Equations (2.15) and (2.16) be reasonable? Clearly, if the independence of potential outcomes, as expressed in Equation (2.6), is valid because the treatment has been randomly assigned, then Assumptions 1 and 2 in Equations (2.15) and (2.16) are implied. But, for observational data analysis, for which random assignment is infeasible, these assumptions would rarely be justified.

Consider the Catholic school example. If one were willing to assume that those who choose to attend Catholic schools do so for completely random reasons, then these two assumptions could be asserted. 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 education with a religious foundation would not be expected to score higher on math and reading tests if educated in Catholic schools rather than public schools. This possibility also seems unlikely, in part because it implies that those with a distaste for education with a religious foundation do not attend Catholic schools. It seems reasonable to assume that these students would perform substantially worse in Catholic schools than the typical students who do attend Catholic schools.

Thus, at least for the Catholic school example, there seems no way to justify the naive estimator as a consistent and unbiased estimator of the ATE. We encourage the reader to consider all of the examples presented in Chapter 1, and we suspect that all will agree that Assumptions 1 and 2 in Equations (2.15) and (2.16) cannot both be sustained for any of them.

Finally, it is important to recognize that assumptions such as these can 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 inconsistent and biased for the ATE, but it is now consistent and unbiased for the ATC. 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 (2.8) as the ATC.

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 (2.7) as the ATT.

Considering the validity of Assumptions 1 and 2 separately shows that the naive estimator may be inconsistent and biased for the ATE and yet may be consistent and unbiased for either the ATT or the ATC. 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 in the counterfactual control state than those in the control group, and (2) Assumption 1 is invalid because those in the control group would not do nearly as well in the counterfactual treatment state as those in the treatment group. Or, stated more simply, we may have good theoretical reason to believe that the treatment is more effective for the treatment group than it would be for the control group. Under this scenario, the naive estimator will deliver a consistent and unbiased estimate of the ATT, even though it is still inconsistent and biased for both the ATC and the unconditional ATE.

Now, return to the case in which neither Assumption 1 nor Assumption 2 is true. If the naive estimator is therefore inconsistent and biased for the typical average causal effects 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. This strategy amounts to conditioning on one or more variables that identify such strata and then asserting that the naive estimator is consistent and unbiased 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 will explain this strategy in great detail in subsequent chapters. Next, we introduce over-time potential outcomes and then extend the framework to many-valued causes.

2.8 Over-Time Potential Outcomes and Causal Effects

Having shown in the last section that the cross-sectional naive estimator will rarely deliver consistent and unbiased estimates of average causal effects of interest when analyzing observational data, it is natural to then wonder whether observing individuals across time and then estimating similar unconditional differences may be more promising. We will take the position in this book that the power of over-time observation is considerable but that it is also too often oversold and misunderstood. In this section, we lay out the basic potential outcome model when observations occur in more

than one time period, moving from the case of a single individual or unit to multiple individuals or units. We reserve our full treatment of the strengths and weaknesses of alternative estimators using repeated observations for Chapter 11.

2.8.1 A Single Unit Over Time

Consider the analysis of a single unit, observed during time intervals indexed by a discrete counter t that increases from 1 to T. The outcome variable is Y_t, which has observed values $\{y_1, y_2, y_3, \ldots, y_T\}$. Suppose that we have a two-state causal variable, D_t, that is equal to 1 if the treatment is in place during a time period t and is equal to 0 otherwise.

Because we are considering only one unit of analysis – possibly an individual but more likely a school, organization, city, state, country, or other aggregate unit – we do not have either a control group or a treatment group. Instead, we have a single unit that is exposed to the treatment state and the control state at different points in time. The fundamental problem/reality of causal inference now is that we cannot observe the same unit at the same time in both the treatment state and the control state.

For an analysis of a single unit, it only makes sense to consider designs where we have at least some pretreatment data and where the unit under consideration spends at least one time period in the treatment state. In particular, we will label the time period in which the treatment is initiated as t^*, and our restriction to situations in which pretreatment data are available requires that $1 < t^* \leq T$. We will allow the treatment to persist for one or more time periods, from t^* through $t^* + k$, where $k \geq 0$. Once the treatment ends, following $t^* + k$, we will not allow the treatment to be reintroduced before the full observation window terminates at T.

We can set up the potential outcome model in the following way to capture the basic features of before-and-after designs for a single unit of analysis:

1. Before the treatment is introduced (for $t < t^*$):[28]

$$D_t = 0$$
$$Y_t = Y_t^0$$

2. While the treatment is in place (from t^* through $t^* + k$):

$$D_t = 1$$
$$Y_t = Y_t^1$$
$$Y_t^0 \text{ exists but is counterfactual}$$

[28] Although in theory counterfactual values Y_t^1 exist in pretreatment time periods, these values are not typically considered. If one were interested in asking what the treatment effect would have been if the treatment had been introduced in an earlier time period, then these counterfactual values would need to be introduced into the analysis.

3. After the treatment ends (for time periods $t > (t^* + k)$):[29]

$$D_t = 0$$
$$Y_t = Y_t^1$$
$$Y_t^0 \text{ exists but is counterfactual.}$$

For a single unit, the causal effect of the treatment is

$$\delta_t = Y_t^1 - Y_t^0, \tag{2.17}$$

and these effects may exist in more than one time period t, depending on the duration of the treatment and whether the treatment is assumed to be a reversible treatment state or a permanent change that cannot be undone. Studies are often unclear on maintained assumptions such as these, as well as on the distinctions between time periods of types 2 and 3. Our setup is very general and can accommodate many alternative types of studies with only minor modifications, including those for which time periods of types 2 or 3 are unobserved.[30] Because such assumptions and design features will always be application-specific, we offer a worked example next.

The Year of the Fire Horse

For a concrete example that reveals the possible power of over-time analysis for a single unit, consider a variant of the demography example on the determinants of fertility introduced in Section 1.3 (see page 17). In addition to the individual-level effects of family background and other life course events on fertility decisions, the causal effects of religion, values, and more general cultural beliefs have been of long-standing interest as well (see Mayer and Marx 1957; Westoff and Jones 1979; Hayford and Morgan 2008; Thornton, Binstock, Yount et al. 2012).

Suppose that the unit of analysis is the birth rate in a single country, estimated from aggregate census data and vital statistics. The example we will consider is presented in Figure 2.1, which displays birth rates in Japan between 1951 and 1980. Following a post-war baby boom, birth rates in Japan were comparatively stable from the late 1950s through the early 1960s. However, in 1966, the birth rate fell precipitously, after which it rebounded in 1967 and then stabilized. From the 1970s onward, Japan's birth rate then resumed its decline, as its population aged and it continued with its demographic transition to a low mortality and low fertility country, as discussed in general in Chapter 1.[31]

[29] Below we will consider an example where $k \leq (T - t^*)$, but we do not mean to imply that the treatment cannot remain in place after the observation window ends at T. In fact, we place no upper bound on values for k. If $(t^* + k) > T$, then none of the time periods of type 3 are observed.

[30] In some applications, the treatment is stipulated to occur between observation intervals. In these cases, time periods of type 2 are assumed to be unobserved. Typically, in this case D_t is assumed to be equal to 1 for at least the first time period of type 3 in order to indicate that the treatment was initiated in the unobserved time periods of type 2. Others studies imply that $t^* + k = T$, such that the treatment is present through the full posttreatment observation window. In these cases, time periods of type 3 are unobserved.

[31] For a comprehensive consideration of the post-1973 "baby bust" in Japan, see Retherford and Ogawa (2006).

One could ask many causal questions about the trend in Japan's birth rate in Figure 2.1, but the natural first question is, What caused the dramatic decline in the birth rate in 1966? The demographic consensus is the following. Every 60 years, two cycles within the Asian zodiac calendar – one over twelve animals and one over five elements – generate a year of the "fire horse." A folk belief exists that families who give birth to babies designated as fire horses will suffer untold miseries, particularly so if the baby is a girl. Enough couples supposedly held this belief in the years around 1966 that they adjusted their fertility behavior accordingly (see Hodge and Ogawa 1991).

In their discussion of causal analysis in demography, Ní Bhrolcháin and Dyson (2007:8) consider this example and write that "demographers naturally interpret this event in a causal way, without worrying about the formalities of causal inference." Although we agree that the assertion of a fire-horse causal effect on Japanese birth rates in 1966 does not require a formal treatment to convince most demographers, we will nonetheless use this example to demonstrate an over-time analysis of a causal effect for a single unit of analysis.

The first issue to consider is measurement of the outcome. The outcome for Figure 2.1 is known as the crude birth rate, which is the number of live births per 1,000 persons alive in the same year. A more refined outcome could be constructed,

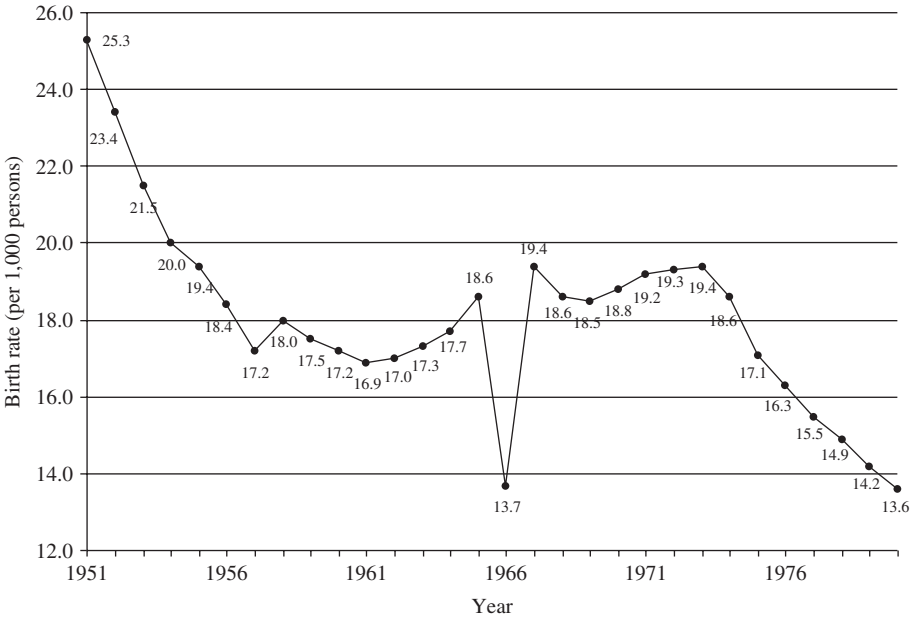

Figure 2.1 Crude birth rates in Japan, 1951–1980.
Source: Table 2-24, Live Births by Sex and Sex Ratio of Live Births, Bureau of Statistics of Japan. Accessed at http://www.stat.go.jp/data/chouki/zuhyou/02-24.xls on March 11, 2013.

standardizing for cohort sizes of women of childbearing age. An analogous drop in standardized birth rates would still be present for 1966.[32]

What are the causal states that generate the seemingly obvious causal effect? At the individual level, the causal states could be quite specific ("believing that having a fire-horse baby is less desirable than having a baby in another year") or considerably more broad ("believing in the relevance of zodiac calendars"). For our purposes here, the particular individual-level causal states do not matter because the causal states are at the national level for this analysis. The states are "fire-horse year in the zodiac calendar" and "not," and they are aligned, not with observed treatment and control groups, but with the observed years, $D_{1966} = 1$ versus $D_{t \neq 1966} = 0$. In the year 1966, the treatment generates two corresponding potential outcome variables, Y^1_{1966} and Y^0_{1966}, which define the effect of interest for the birth rate in Japan in 1966:

$$\delta_{1966} = y^1_{1966} - y^0_{1966}.$$

We can estimate this effect as

$$\begin{aligned} \hat{\delta}_{1966} &= y_{1966} - \hat{y}^0_{1966} \\ &= 13.7 - \hat{y}^0_{1966}. \end{aligned}$$

The estimate $\hat{\delta}_{1966}$ is the observed birth rate in 1966, 13.7, minus an estimate of the counterfactual birth rate, $\hat{y}^0_{1966}$, that would have been observed if 1966 had not been the year of the fire horse.

What is a reasonable estimated value for the true counterfactual outcome, y^0_{1966}? How about 18.6, which is the birth rate for 1965? Ní Bhrolcháin and Dyson (2007:30) reason that "some births that might have taken place in 1966 were transferred – either in fact, or via the year of occurrence reported by the parents at the time of registration – to the adjacent years." Such a possibility is evident in 1965, where the birth rate appears elevated in comparison to prior years. In fact, the birth rate appears to be slightly higher in 1964 as well, relative to the prevailing trend in the early 1960s. What about the value for 1967, 19.4? Again, the birth rate appears to be higher here too, and possibly again in 1968. Following Ní Bhrolcháin and Dyson's reasoning (and also Hodge and Ogawa 1991), these higher rates in 1964, 1965, 1967, and 1968 could be present because parents were especially eager to have children before or after the year of the fire horse and adjusted their behavior accordingly. Given the uncertainties of conception, at least some of these parents started their avoidance early or failed to conceive a child until after the year of the fire horse.

In our experience discussing this example with other researchers, most will settle on the average of the two values for 1963 and 1969, $(17.3 + 18.5)/2 = 17.9$, as a reasonable estimate of the counterfactual value, y^0_{1966}. The result of such a choice is an estimated causal effect,

$$\hat{\delta}_{1966} = 13.7 - 17.9 = -4.2,$$

[32] At least some of the overall downward trend would disappear because the trend in Figure 2.1 is produced in part by the aging of the population (which itself is a function of the return to peace following World War II and continuing declines in mortality). Hodge and Ogawa (1991) document these trends and consider alternative adjustments for other demographic trajectories.

that implies a decline in the crude birth rate of 31 percent. At the individual level, and ignoring rates of twins and so on, this estimate suggests that nearly 1 out of 3 mothers who would have given birth in 1966 did not do so because 1966 was the year of the fire horse. Notice also that, in selecting the average of the birth rates in 1963 and 1969 as the most reasonable estimate of the counterfactual value, one is thereby assuming positive fire-horse-year effects in 1964, 1965, 1967, and 1968, which were non–fire-horse years. As a result, in order to estimate the overall effect of the year of the fire horse on the population structure of Japan, as determined by the year-by-year evolution of what is known as the total fertility rate, one would need to model and then appropriately combine five year-specific effects, δ_{1964} through δ_{1968}, based on additional corresponding treatment states for the specific years.

Now consider what has been learned and what has not. Visual inspection of Figure 2.1 is probably convincing on its own that something causal happened in 1966 that altered birth rates in Japan. However, the specific explanation that is accepted in demography is based on substantive knowledge of Asian zodiac calendars, as well as cultural beliefs based upon them that were sufficiently widespread in 1966. The over-time design did not generate this knowledge, even though it provided the incentive to uncover it.

Has anything deeper been learned from this effect? Certainly the effect supports the more general conclusion that cultural beliefs have been a cause of fertility decisions in Japan in the past. This conclusion may then further bolster the overall perspective in empirical demography that fertility decisions across the world are likely shaped by cultural beliefs and cannot be reduced to a cold rational calculus of the direct costs and psychic benefits of producing offspring.

With a little extra work, we have been able to generate the reasonable estimate that the effect in 1966 was a reduction of the birth rate of 31 percent, and we also took the position that there were very likely positive near-fire-horse-year effects in the two years on either side of 1966. Notwithstanding these successes, our analysis yields no information on which types of couples changed their fertility behavior. The relevance of the zodiac calendar surely varies across couples in nonrandom ways, and such patterns would be useful to know in order to develop additional perspective on the broader consequences of the effect. We have also not learned how many women, or which women, had fewer children or instead simply had children who were one or two years younger or older than they would have been if their children had been born in 1966.

We will return to this type of example in Chapter 11, where we will consider the deeper modeling issues that accompany the estimation of these types of effects. Often referred to as interrupted time series (ITS) designs, there are formal time series analysis procedures for generating best estimates of counterfactual values, and these may be easier to defend than our ad hoc choice of 17.9 as the most reasonable estimate of the crucial counterfactual value that determines the size of the casual effect. As we will discuss in detail in Chapter 11, the main weakness of these designs is their rarity. In the observational social sciences, few examples are as clear-cut as the year of the fire horse. More commonly, the causal shocks to outcomes unfolding over time are less dramatic, and, as a result, they are more difficult to separate from underlying trends.

Notice also that we have chosen in this section an aggregate unit – a country – as our example of the analysis of a causal effect for a "single unit." In part, this decision reflects our position on what can be learned by studying individuals in isolation. Surely there are genuine individual-level causal effects, some of which can be discerned from examining the lives of particular individuals over time. The challenge is what in general can be learned from documenting such apparent effects, and whether analyses can be strengthened by considering individuals in groups, especially in representative samples. This is the focus of our next section, where we introduce the potential outcome model for many units over time.

2.8.2 Many Units Over Time

Suppose now that we have a collection of individuals observed at multiple points in time. Generally referred to as panel data or longitudinal data, outcomes and causes now vary over both individuals and time. Consider two examples. First, researchers who study the charter school effect often have access to samples of students observed over multiple grades in both charter schools and regular public schools. Second, researchers who study the effects of father absence that results from divorce typically have access to data from random samples of families. These data often include measures of child development outcomes before and after the divorce that triggers father absence.

We will now extend the potential outcome framework to consider the estimation of causal effects with such over-time data on samples of individuals. In earlier sections of this chapter, we dropped subscripting on i for brevity when discussing causal effects. For this section, in which we must deal now with some quantities that vary only over individuals, others that vary only over time, and others that vary over both, we subscript with i for individuals and with t for time. For a two-state cause, the potential outcome variables are Y_{it}^1, Y_{it}^0, and the observed variables are D_{it} and Y_{it}.[33] As in the last section, we will allow the observation window to run from $t = 1$ to T. Unlike the last section, we will not utilize notation for a focal time interval, t^*, in which the treatment is introduced. Here, we want to preserve the possibility that the treatment is introduced at different times for different individuals.

We will also now distinguish between two different treatment indicator variables. D_{it} is a time-varying variable that indicates whether individual i receives the treatment in time period t. In contrast, D_i^* is a time-constant variable that indicates whether individual i ever receives the treatment at any point in time during the observation window of the study (i.e., in any time period from $t = 1$ to T). D_{it} is best thought of as a *treatment exposure indicator*, and D_i^* is best thought of as a *treatment group indicator.*

The setup for the potential outcome model with panel data follows directly from these definitions. For members of the control group, $Y_{it} = Y_{it}^0$ for all time periods t. For members of the treatment group, $Y_{it} = Y_{it}^0$ before treatment exposure, and $Y_{it} = Y_{it}^1$

[33] In some cases, the subscripting is redundant. For example, in Sections 2.4 and 2.7, we represented the causal effect as δ, recognizing that this effect can vary over individuals. In this section, we will represent the individual-level causal effect always as δ_i, so that it is clear that in this form we are assuming that it does not vary with time. For a time-varying causal effect, we would instead need to subscript it as δ_{it}.

after treatment exposure begins. Altogether, Y_{it} is defined with reference to D_{it}, such that

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0, \tag{2.18}$$

where D_{it} remains equal to 0 over time for all members of the control group ($D_i^* = 0$) but varies between 0 and 1 for all members of the treatment group ($D_i^* = 1$).

Consider a concrete application of this notation using both the treatment exposure indicator and the treatment group indicator. If no one is exposed to the treatment in time period $t = 1$ or $t = 2$, then $D_{i1} = 0$, $D_{i2} = 0$, $Y_{i1} = Y_{i1}^0$, and $Y_{i2} = Y_{i2}^0$ for those in the control group ($D_i^* = 0$) and for those in the treatment group ($D_i^* = 1$). But, if the treatment is then introduced to all members of the treatment group in time period $t = 3$, then the values of D_{it} and Y_{it} diverge across the treatment and control groups in time period 3. Now, for those in the control group ($D_i^* = 0$), $D_{i3} = 0$ and $Y_{i3} = Y_{i3}^0$. But, for those in the treatment group ($D_i^* = 1$), $D_{i3} = 1$ and $Y_{i3} = Y_{i3}^1$.

The distinction between D_{it} and D_i^* reveals the potential value of panel data. For a cross-sectional study in which all observation occurs in a single time period, $D_{it} = D_i^*$. However, a panel dataset over multiple time periods allows a researcher to consider how treatment exposure (D_{it}) can be separated from treatment group membership (D_i^*) and then exploit this difference to estimate the causal effect. Consider two types of analysis.

First, when individuals receive the treatment at different times or do not receive the treatment at all, it is possible to observe how Y_{it}^0 changes over time for some individuals after others have received the treatment. As a result, it may be possible to make reasonable predictions about how Y_{it}^0 would have evolved over time for individuals in the treatment group during the posttreatment time period. If predictions of these counterfactual trajectories are reasonable, inferences about the causal effect of the treatment, δ_{it}, in time period t, can be made by comparison of the observed Y_{it} in time period t for those who are treated ($D_i^* = 1$) with predictions of their corresponding counterfactual values of Y_{it}^0 in time period t. The crux of the matter, of course, is how to use observed values of $Y_{it} = Y_{it}^0$ in time period t among those in the control group ($D_i^* = 0$) to make reasonable predictions about posttreatment counterfactual values of Y_{it}^0 for those in the treatment group ($D_i^* = 1$).

Second, in situations where it is reasonable to assume that $E[Y_{it}^0]$ is evolving in parallel fashion across the treatment and control groups and that treatment assignment is unrelated to the values of the outcome prior to the treatment, it is possible to estimate the ATT by offering a model that (1) calculates the average difference in the outcome in the treatment group between the pretreatment and posttreatment time periods and (2) subtracts from this difference the average difference in the outcome for the control group between the same two time periods.[34] In this case, the most important issue to consider is whether it is reasonable to assume parallel trajectories and that treatment assignment is unrelated to pretreatment values of the outcome. The latter assumption would be unreasonable if individuals with comparatively low or comparatively high values of the pretreatment outcome, net of other determinants

[34] Moreover, if it is reasonable to assume that self-selection on the causal effect is absent, then the ATT is equal to the ATE and ATC. A consistent and unbiased estimate of the ATT from a difference-based estimator is then also consistent and unbiased for both the ATE and ATC.

of the outcome, select into the treatment, either under the assumption that they are especially suited to the treatment because of their recent strong performance or that they are especially in need of the treatment because of their recent weak performance.

We will discuss a variety of panel data estimation strategies in Chapter 11, but we want to foreshadow two basic conclusions here to temper the optimism that many may feel after considering our prior two paragraphs. First, panel data estimators based only on posttreatment observations do not usually improve on the cross-sectional estimators we will present in this book. In our experience, analysts are often too sanguine about the clarifying power of observing the evolution of outcome variables for those who are always observed under exposure to the treatment of interest (e.g., students always enrolled in Catholic high schools relative to students always enrolled in public high schools, with no data on either group available prior to high school). Second, one needs data from multiple pretreatment time periods and/or very well-developed theory to justify required assumptions before the gains to panel data are clear, especially given the other ancillary patterns, such as panel attrition, that must also often be modeled.

2.9 The Potential Outcome Model for Many-Valued Treatments

So far in this chapter, we have focused our presentation of the potential outcome model on binary causal variables, conceptualized as dichotomous variables that indicate whether individuals are observed in treatment and control states. As we show in this section, the counterfactual framework can be used to analyze causal variables with more than two categories.

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-valued treatments presented in Table 2.4 (which is simply an expansion of Table 2.1 to 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

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

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

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 may not 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.[35] 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 easily 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 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.

[35] 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. We will discuss the importance of considering such groups of "always takers" and "never takers" when we present instrumental variable estimators in Chapter 9.

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

Now consider a variant on the education-earnings example. 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 diploma (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 diploma.

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 are 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 diploma 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 diplomas. 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 diplomas.

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 diplomas: $E[Y^{\text{BA}} - Y^{\text{HS}}|\text{HS} = 1]$. In this case, the causal effect of attaining a bachelor's degree for those who typically attain degrees other than a high school diploma 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 expectations. 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 farthest from the direct comparison one hopes to estimate. We will discuss these sorts of assumptions in Chapter 12.

Aside from the expansion of the number of causal states, potential outcomes, and treatment effects, all other features of the potential outcome model remain essentially the same. SUTVA is typically still 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 inconsistency and 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 discussing the estimation of treatment effects using linear regression for many-valued treatments in Section 6.6.1, we will consider 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.

2.10 Conclusions

In this chapter, we have introduced the main components of the potential outcome model, which is a foundational piece of the counterfactual model of causality for observational research. We defined individual-level causal effects as the what-if differences in potential outcomes that would result from being exposed to alternative causal states. We then presented the assumption of causal effect stability – the stable unit treatment value assumption – that is frequently relied on when estimating effects defined by potential outcomes. We defined average causal effects at the population level, considered how ineffective the simple mean-difference estimator is for estimating average causal effects with observational data, and concluded with extensions of the potential

outcome model for effects observed over time and for effects defined across many values of the cause. In the next chapter, we introduce the directed graph approach to causal analysis, which we see as the second foundational piece of the counterfactual model of causality for observational research.

2.11 Appendix to Chapter 2: 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 5 through 7.

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, 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 4 and 5). 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 with good fortune 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. [Many researchers do not have such good fortune and instead must analyze a dataset with measures of only a subset of the variables in S, which we will typically label X. These researchers have access to a dataset $\{y_i, d_i, x_i\}_{i=1}^N$ and model $\Pr_N(Y, D, X)$, which does not converge to $\Pr(Y, D, S)$.]

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 a country's population older than 18) and makes it cumbersome to discuss some of the estimators we will consider (e.g., in Chapter 5, 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 often requires constraints, such as homogeneity of treatment effects, that are too 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, 2010) 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.