CHAPTER 1

Causality: The Basic Framework

1.1 INTRODUCTION

In this introductory chapter we set out our basic framework for causal inference. We discuss three key notions underlying our approach. The first notion is that of *potential outcomes*, each corresponding to one of the levels of a *treatment* or *manipulation*, following the dictum "no causation without manipulation" (Rubin, 1975, p. 238). Each of these potential outcomes is *a priori* observable, in the sense that it could be observed if the unit were to receive the corresponding treatment level. But, *a posteriori*, that is, once a treatment is applied, at most one potential outcome can be observed. Second, we discuss the necessity, when drawing causal inferences, of observing *multiple units*, and the utility of the related *stability* assumption, which we use throughout most of this book to exploit the presence of multiple units. Finally, we discuss the central role of the *assignment mechanism*, which is crucial for inferring causal effects, and which serves as the organizing principle for this book.

1.2 POTENTIAL OUTCOMES

In everyday life, causal language is widely used in an informal way. One might say: "My headache went away because I took an aspirin," or "She got a good job last year because she went to college," or "She has long hair because she is a girl." Such comments are typically informed by observations on past exposures, for example, of headache outcomes after taking aspirin or not, or of characteristics of jobs of people with or without college educations, or the typical hair length of boys and girls. As such, these observations generally involve informal statistical analyses, drawing conclusions from associations between measurements of different quantities that vary from individual to individual, commonly called *variables* or *random variables* – language apparently first used by Yule (1897). Nevertheless, statistical theory has been relatively silent on questions of causality. Many, especially older, textbooks avoid any mention of the term other than in settings of randomized experiments. Some mention it mainly to stress that correlation or association is not the same as causation, and some even caution their readers to avoid

using causal language in statistics. Nevertheless, for many users of statistical methods, causal statements are exactly what they seek.

The fundamental notion underlying our approach is that causality is tied to an *action* (or manipulation, treatment, or intervention), applied to a *unit*. A unit here can be a physical object, a firm, an individual person, or collection of objects or persons, such as a classroom or a market, at a particular point in time. For our purposes, the same physical object or person at a different time is a different unit. From this perspective, a causal statement presumes that, although a unit was (at a particular point in time) subject to, or exposed to, a particular action, treatment, or regime, the same unit could have been exposed to an alternative action, treatment, or regime (at the same point in time). For instance, when deciding to take an aspirin to relieve your headache, you could also have choosen not to take the aspirin, or you could have chosen to take an alternative medicine. In this framework, articulating with precision the nature and timing of the action sometimes requires a certain amount of imagination. For example, if we define race solely in terms of skin color, the action might be a pill that alters only skin color. Such a pill may not currently exist (but, then, neither did surgical procedures for heart transplants hundreds of years ago), but we can still imagine such an action.

This book primarily considers settings with two actions, although many of the extensions to multi-valued treatments are conceptually straightforward. Often one of these actions corresponds to a more active treatment (e.g., taking an aspirin) in contrast to a more passive action (e.g., not taking the aspirin). In such cases we sometimes refer to the first action as the *active treatment* as opposed to the *control treatment*, but these are merely labels and formally the two treatments are viewed symmetrically. In some cases, when it is clear from the context, we refer to the more active treatment simply as the "treatment" and the other treatment as the "control."

Given a unit and a set of actions, we associate each action-unit pair with a *potential outcome*. We refer to these outcomes as potential outcomes because only one will ultimately be realized and therefore possibly observed: the potential outcome corresponding to the action actually taken. *Ex post*, the other potential outcomes cannot be observed because the corresponding actions that would lead to them being realized were not taken. The causal effect of one action or treatment relative to another involves the comparison of these potential outcomes, one realized (and perhaps, though not necessarily, observed), and the others not realized and therefore not observable. Any treatment must occur temporally before the observation of any associated potential outcome is possible.

Although the preceding argument may appear obvious, its force is revealed by its ability to clarify otherwise murky concepts, as can be demonstrated by considering the three examples of informal "because" statements presented in the first paragraph of this section. In the first example, it is clear what the action is: I took an aspirin, but at the time that I took the aspirin, I could have followed the alternate course of not taking an aspirin. In that case, a different outcome might have resulted, and the "because" statement is causal in the perspective taken in this book as it reflects the comparison of those two potential outcomes. In the second example, it is less clear what the treatment and its alternative are: she went to college, and at the point in time when she decided to go to college, she could have decided not to go to college. In that case, she might have had a different job a year ago, and the implied causal statement compares the quality of the job she actually had then to the quality of the job she would have had a year ago, had she not

gone to college. However, in this example, the alternative treatment is somewhat murky: had she not enrolled in college, would she have enrolled in the military, or would she have joined an artist's colony? As a result, the potential outcome under the alternative action, the job obtained a year ago without enrolling in college, is not as well defined as in the first example.

In the third example, the alternative action is not at all clear. The informal statement is "she has long hair because she is a girl." In some sense the implicit treatment is being a girl, and the implicit alternative is being a boy, but there is no action articulated that would have made her a boy and allowed us to observe the alternate potential outcome of hair length for this person as a boy. We could clarify the causal effect by defining such an action in terms of surgical procedures, or hormone treatments, all with various ages at which the action to be taken is specified, but clearly the causal effect is likely to depend on the particular alternative action and timing being specified. As stated, however, there is no clear action described that would have allowed us to observe the unit exposed to the alternative treatment. Hence, in our approach, this "because" statement is ill-defined as a causal statement.

It may seem restrictive to exclude from consideration such causal questions. However, the reason to do so in our framework is that without further explication of the intervention being considered, the causal question is not well defined. One can make many of these questions well posed in our framework by explicitly articulating the alternative intervention. For example, if the question concerns the causal effect of "race," then an ethnicity change on a *curriculum vitae* (or its perception, as in Bertrand and Mullainathan, 2004) defines one causal effect being contemplated, whereas if the question concerns a futuristic "at conception change of chromosomes determining skin color," there is a different causal effect being contemplated. With either manipulation, the explicit description of the intervention makes the question a plausible causal one in our framework.

A closely related way of interpreting the qualitative difference between the three "causal" statements is to consider, after application of the actual treatment, the counterfactual value of the potential outcome corresponding to the treatment not applied. In the first statement, the treatment applied is "aspirin taken," and the counterfactual potential outcome is the state of your headache under "aspirin not taken"; here it appears unambiguous to consider the counterfactual outcome. In the second example, the counterfactual outcome is her job a year ago had she decided not to go to college, which is not as well defined. In the last example, the counterfactual outcome – the person's hair length if she were a boy rather than a girl (note the lack of an action in this statement) – is not at all well defined, and therefore the causal statement is correspondingly poorly defined. In practice, the distinction between well and poorly defined causal statements is one of degree. The important point is, however, that causal statements become more clearly defined by more precisely articulating the intervention that would have made the alternative potential outcome the realized one.

1.3 DEFINITION OF CAUSAL EFFECTS

Let us consider the case of a single unit, I, at a particular point in time, contemplating whether or not to take an aspirin for my headache. That is, there are two treatment levels,

Table 1.1. ***Example of Potential Outcomes and Causal Effect with One Unit***

Unit	Potential Outcomes		Causal Effect
	Y(Aspirin)	Y(No Aspirin)	
You	No Headache	Headache	Improvement due to Aspirin

taking an aspirin, and not taking an aspirin. If I take the aspirin, my headache may be gone, or it may remain, say, an hour later; we denote this outcome, which can be either "Headache" or "No Headache," by Y(Aspirin). (We could use a finer measure of the status of my headache an hour later, for example, rating my headache on a ten-point scale, but that does not alter the fundamental issues involved here.) Similarly, if I do not take the aspirin, my headache may remain an hour later, or it may not; we denote this potential outcome by Y(No Aspirin), which also can be either "Headache," or "No Headache." There are therefore two potential outcomes, Y(Aspirin) and Y(No Aspirin), one for each level of the treatment. The causal effect of the treatment involves the comparison of these two potential outcomes.

Because in this example each potential outcome can take on only two values, the unit-level causal effect – the comparison of these two outcomes for the same unit – involves one of four (two by two) possibilities:

1. Headache gone only with aspirin:
   Y(Aspirin) = No Headache, Y(No Aspirin) = Headache
2. No effect of aspirin, with a headache in both cases:
   Y(Aspirin) = Headache, Y(No Aspirin) = Headache
3. No effect of aspirin, with the headache gone in both cases:
   Y(Aspirin) = No Headache, Y(No Aspirin) = No Headache
4. Headache gone only without aspirin:
   Y(Aspirin) = Headache, Y(No Aspirin) = No Headache

Table 1.1 illustrates this situation assuming the values Y(Aspirin) = No Headache, Y(No Aspirin) = Headache. There is a zero causal effect of taking aspirin in the second and third possibilities. In the other two cases the aspirin has a causal effect, making the headache go away in one case and not allowing it to go away in the other.

There are two important aspects of this definition of a causal effect. First, the definition of the causal effect depends on the potential outcomes, but it does *not* depend on which outcome is actually observed. Specifically, whether I take an aspirin (and am therefore unable to observe the state of my headache with no aspirin) or do not take an aspirin (and am thus unable to observe the outcome with an aspirin) does not affect the definition of the causal effect. Second, the causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment. In particular, the causal effect is *not* defined in terms of comparisons of outcomes at different times, as in a before-and-after comparison of my headache before and after deciding to take or not to take the aspirin. "The fundamental problem of causal inference" (Holland, 1986, p. 947) is therefore the problem that at most one of the potential outcomes can be realized and thus observed. If the action you take is Aspirin, you observe Y(Aspirin) and

Table 1.2. ***Example of Potential Outcomes, Causal Effect, Actual Treatment, and Observed Outcome with One Unit***

Unit	Not Observable			Known	
	Potential Outcomes		Causal Effect	Actual Treatment	Observed Outcome
	Y(Aspirin)	Y(No Aspirin)			
You	No Headache	Headache	Improvement due to Aspirin	Aspirin	No Headache

will never know the value of Y(No Aspirin) because you cannot go back in time. Similarly, if your action is No Aspirin, you observe Y(No Aspirin) but cannot know the value of Y(Aspirin). Likewise, for the college example, we know the outcome given college attendance because the woman actually went to college, but we will never know what job she would have had if she had not gone to college. In general, therefore, even though the unit-level causal effect (the comparison of the two potential outcomes) may be well defined, by definition we cannot learn its value from just the single realized potential outcome. Table 1.2 illustrates this concept for the aspirin example, assuming the action taken was that you took the aspirin.

For the *estimation* of causal effects, as opposed to the *definition* of causal effects, we will need to make different comparisons from the comparisons made for their definitions. For estimation and inference, we need to compare *observed* outcomes, that is, observed realizations of potential outcomes, and because there is only one realized potential outcome per unit, we will need to consider multiple units. For example, a before-and-after comparison of the same physical object involves distinct units in our framework, and also the comparison of two different physical objects at the same time involves distinct units. Such comparisons are critical for *estimating* causal effects, but they do not *define* causal effects in our approach. For estimation it will also be critical to know about, or make assumptions about, the reason why certain potential outcomes were realized and not others. That is, we will need to think about the *assignment mechanism*, which we introduce in Section 1.7. However, we do not need to think about the assignment mechanism for defining causal effects: we merely need to do the thought experiment of the manipulations leading to the definition of the potential outcomes.

1.4 CAUSAL EFFECTS IN COMMON USAGE

The definition of a causal effect given in the previous section may appear a bit formal, and the discussion a bit ponderous, but the presentation is simply intended to capture the way we use the concept in everyday life. Also, implicitly this definition of causal effect as the comparison of potential outcomes is frequently used in contemporary culture, for example, in the movies. Many of us have seen the movie *It's a Wonderful Life*, with Jimmy Stewart as George Bailey. In this movie George Bailey becomes very depressed and states that the world would have been a better place had he never been born. At the appropriate moment an angel appears and shows him what the world would have been like had he not been born. The actual world is the real, observed outcome, but the

angel shows George the other potential outcome, had George not been born. Not only are there obvious consequences, like his own children not existing, but there are many other untoward events. For example, his younger brother, who was in actual life a World War II hero, in the counterfactual world drowns in a skating accident at age eight because George was not there to save him. In the counterfactual world a pharmacist fills in a wrong prescription and is convicted of manslaughter because George was not there to catch the error as he did in the actual world. The causal effect of George not being born is the comparison of the entire stream of events in the actual world with George in it, with the entire stream of events in the counterfactual world without George in it. In reality we would never be able to see both worlds, but in the movie George gets to observe both.

Another interesting comparison is to the "but-for" concept in legal settings. Suppose someone committed an action that is harmful, and a second person suffered damages. From a legal perspective, the damage that the second person is entitled to collect is the difference between the economic position of the plaintiff had the harmful event not occurred (the economic position "but-for" the harmful action) and the actual economic position of the plaintiff. Clearly, this is a comparison of the potential outcome that was not realized and the realized potential outcome, this difference being the causal effect of the harmful action.

1.5 LEARNING ABOUT CAUSAL EFFECTS: MULTIPLE UNITS

Although the *definition* of causal effects does not require more than one unit, *learning* about causal effects typically requires multiple units. Because with a single unit we can at most observe a single potential outcome, we must rely on multiple units to make causal inferences. More specifically, we must observe multiple units, some exposed to the active treatment, some exposed to the alternative (control) treatment.

One option is to observe the same physical object under different treatment levels at different points in time. This type of data set is a common source for personal, informal assessments of causal effects. For example, I might feel confident that an aspirin is going to relieve my headache within an hour, based on previous experiences, including episodes when my headache went away when I took an aspirin, and episodes when my headache did not go away when I did not take aspirin. In that situation, my views are shaped by comparisons of multiple units: myself at different times, taking and not taking aspirin. There is sometimes a tendency to view the same physical object at different times as the same unit. We view this as a fundamental mistake. The same physical unit, "myself at different times," is not the same unit in our approach to causality. Time matters for many reasons. For example, I may become more or less sensitive to aspirin, evenings may differ from mornings, or the initial intensity of my headache may affect the result. It is often reasonable to assume that time makes little difference for inanimate objects – we may feel confident, from past experience, that turning on a faucet will cause water to flow from that tap – but this assumption is typically less reasonable with human subjects, and it is never correct to confuse assumptions (e.g., about similarities between different units), with definitions (e.g., of a unit, or of a causal effect).

As an alternative to observing the same physical object repeatedly, one might observe different physical objects at approximately the same time. This situation is another common source for informal assessments of causal effects. For example, if both you

and I have headaches, but only one of us takes an aspirin, we may attempt to infer the efficacy of taking aspirin by comparing our subsequent headaches. It is more obvious here that "you" and "I" at the same point in time are different units. Your headache status after taking an aspirin can obviously differ from what my headache status would have been had I taken an aspirin. I may be more or less sensitive to aspirin, or I may have started with a more or less severe headache. This type of comparison, often involving many different individuals, is widely used in informal assessments of causal effects, but it is also the basis for many formal studies of causal effects in the social and biomedical sciences. For example, many people view a college education as economically beneficial to future career outcomes based on comparisons of the careers of individuals with, and individuals without, college educations.

By itself, however, the presence of multiple units does not solve the problem of causal inference. Consider the aspirin example with two units, You and I, and two possible treatments for each unit, aspirin or no aspirin. For simplicity, assume that the two available aspirin tablets are equally effective. There are now a total of four treatment levels: you take an aspirin and I do not, I take an aspirin and you do not, we both take an aspirin, or neither of us does. There are therefore four potential outcomes for each of us. For "I" these four potential outcomes are the state of my headache (*i*) if neither of us takes an aspirin, (*ii*) if I take an aspirin and you do not, (*iii*) if you take an aspirin and I do not, and (*iv*) if both of us take an aspirin. "You," of course, have the corresponding set of four potential outcomes. We can still only observe at most one of these four potential outcomes for each unit, namely the one realized corresponding to whether you and I took, or did not take, an aspirin. Thus each level of the treatment now indicates both whether you take an aspirin and whether I do. In this situation, there are six different comparisons defining causal effects for each of us, depending on which two of the four potential outcomes for each unit are conceptually compared $\left(6 = \binom{4}{2}\right)$. For example, we can compare the status of my headache if we both take aspirin with the status of my headache if neither of us takes an aspirin, or we can compare the status of my headache if only you take an aspirin to the status of my headache if we both do.

Although we typically make the assumption that whether you take an aspirin does not affect my headache status, it is important to understand the force of such an assumption. One should not lose sight of the fact that it is an assumption, often a strong and controversial one, not a fact, and therefore may be false. Consider a setting where I take aspirin, and I will have a headache if you do not take an aspirin, whereas I will not have a headache if you do take an aspirin: we are in the same room, and unless you take an aspirin to ease your own headache, your incessant complaining will maintain my headache! Such interactions or spillover effects are an important feature of many educational programs, and often motivate changing the unit of analysis from individual children to schools or other groups of individuals.

1.6 THE STABLE UNIT TREATMENT VALUE ASSUMPTION

In many situations it may be reasonable to assume that treatments applied to one unit do not affect the outcome for another unit. For example, if we are in different locations and have no contact with each other, it would appear reasonable to assume that whether

you take an aspirin has no effect on the status of my headache. (But, as the example in the previous section illustrates, this assumption need not hold if we are in the same location, and your behavior, itself affected by whether you take an aspirin, may affect the status of my headache, or if we communicate by extrasensory perception.) The stable unit treatment value assumption, or SUTVA (Rubin, 1980a) incorporates both this idea that units do not interfere with one another and the concept that for each unit there is only a single version of each treatment level (ruling out, in this case, that a particular individual could take aspirin tablets of varying efficacy):

Assumption 1.1 (SUTVA)
The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

These two elements of the stability assumption enable us to exploit the presence of multiple units for estimating causal effects.

SUTVA is the first of a number of assumptions discussed in this book that are referred to generally as *exclusion restrictions*: assumptions that rely on external, substantive, information to rule out the existence of a causal effect of a particular treatment relative to an alternative. For instance, in the aspirin example, in order to help make an assessment of the causal effect of aspirin on headaches, we could exclude the possibility that your taking or not taking aspirin has any effect on my headache. Similarly, we could exclude the possibility that the aspirin tablets available to me are of different strengths. Note, however, that these assumptions, and other restrictions discussed later, are not directly informed by observations – they are assumptions. That is, they rely on previously acquired knowledge of the subject matter for their justification. Causal inference is generally impossible without such assumptions, and thus it is critical to be explicit about their content and their justifications.

1.6.1 SUTVA: No Interference

Consider, first, the no-interference component of SUTVA – the assumption that the treatment applied to one unit does not affect the outcome for other units. Researchers have long been aware of the importance of this concept. For example, when studying the effect of different types of fertilizers in agricultural experiments on plot yields, traditionally researchers have taken care to separate plots using "guard rows," unfertilized strips of land between fertilized areas. By controlling the leaching of different fertilizers across experimental plots, these guard rows make SUTVA more credible; without them we might suspect that the fertilizer applied to one plot affected the yields in contiguous plots.

In our headache example, in order to address the no-interference assumption, one has to argue, on the basis of a prior knowledge of medicine and physiology, that someone else taking an aspirin in a different location cannot have an effect on my headache. You might think that we could learn about the magnitude of such interference from a separate experiment. Suppose people are paired, with each pair placed in a separate room. In each pair one randomly choosen individual is selected to be the "designated treated" individual and the other the "designated control" individual. Half the pairs are then randomly

selected to be the "treatment pairs" and the other half selected to be "control pairs," with the "designated treated" individual in the treatment pairs given aspirin and the "designated treated" individual in the control pairs given a placebo. The outcome would then be the status of the headache of the "control" person in each pair. Although such an experiment could shed some light on the plausibility of our no-interference assumption, this experiment relies itself on a more distant version of SUTVA – that treatments assigned to one pair do not affect the results for other pairs. As this example reveals, in order to make any assessment of causal effects, the researcher has to rely on assumed existing knowledge of the current subject matter to assert that some treatments do not affect outcomes for some units.

There exist settings, moreover, in which the no-interference part of SUTVA is controversial. In large-scale job training programs, for example, the outcomes for one individual may well be affected by the number of people trained when that number is sufficiently large to create increased competition for certain jobs. In an extreme example, the effect on your future earnings of going to a graduate program in statistics would surely be very different if everybody your age also went to a graduate program in statistics. Economists refer to this concept as a *general equilibrium* effect, in contrast to a *partial equilibrium* effect, which is the effect on your earnings of a statistics graduate degree under the *ceteris paribus* assumption that "everything else" stayed equal. Another classic example of interference between units arises in settings with immunizations against infectious diseases. The causal effect of your immunization versus no immunization will surely depend on the immunization of others: if everybody else is already immunized with a perfect vaccine, and others can therefore neither get the disease nor transmit it, your immunization is superfluous. However, if no one else is immunized, your treatment (immunization with a perfect vaccine) would be effective relative to no immunization. In such cases, sometimes a more restrictive form of SUTVA can be considered by defining the unit to be the community within which individuals interact, for example, schools in educational settings, or specifically limiting the number of units assigned to a particular treatment.

1.6.2 SUTVA: No Hidden Variations of Treatments

The second component of SUTVA requires that an individual receiving a specific treatment level cannot receive different forms of that treatment. Consider again our assessment of the causal effect of aspirin on headaches. For the potential outcome with both of us taking aspirin, we obviously need more than one aspirin tablet. Suppose, however, that one of the tablets is old and no longer contains a fully effective dose, whereas the other is new and at full strength. In that case, each of us may have three treatments available: no aspirin, the ineffective tablet, and the effective tablet. There are thus two forms of the active treatment, both nominally labeled "aspirin": aspirin+ and aspirin−. Even with no interference we can now think of there being three potential outcomes for each of us, the no aspirin outcome Y_i(No Aspirin), the weak aspirin outcome Y_i(Aspirin−) and the strong aspirin outcome Y_i(Aspirin+), with i indexing "I" or "You." The second part of SUTVA either requires that the two aspirin outcomes are identical: Y_i(Aspirin+) $=$ Y_i(Aspirin−), or that I can only get Aspirin+ and you can only get Aspirin− (or *vice versa*). Alternatively we can redefine the treatment as taking

a randomly selected aspirin (either Aspirin− or Aspirin+). In that case SUTVA might be satisfied for the redefined stochastic treatment.

Another example of variation in the treatment that is ruled out by SUTVA occurs when differences in the method of administering the treatment matter. The effect of taking a drug for a particular individual may differ depending on whether the individual was assigned to receive it or chose to take it. For example, taking it after being given the choice may lead the individual to take actions that differ from those that would be taken if the individual had no choice in the taking of the drug.

Fundamentally, the second component of SUTVA is again an exclusion restriction. The requirement is that the label of the aspirin tablet, or the nature of the administration of the treatment, cannot alter the potential outcome for any unit. This assumption does *not* require that all forms of each level of the treatment are identical across all units, but only that unit i exposed to treatment level w specifies a well-defined potential outcome, $Y_i(w)$, for all i and w. One strategy to make SUTVA more plausible relies on redefining the represented treatment levels to comprise a larger set of treatments, for example, Aspirin−, Aspirin+, and no-aspirin instead of only Aspirin and no-aspirin. A second strategy involves coarsening the outcome; for example, SUTVA may be more plausible if the outcome is defined to be dead or alive rather than to be a detailed measurement of health status. The point is that SUTVA implies that the potential outcomes for each unit and each treatment are well-defined functions (possibly with stochastic images) of the unit index and the treatment.

1.6.3 Alternatives to SUTVA

To summarize the previous discussion, assessing the causal effect of a binary treatment requires observing more than a single unit, because we must have observations of potential outcomes under both treatments: those associated with the receipt of the treatment on some units and those associated with no receipt of it on some other units. However, with more than one unit, we face two immediate complications. First, there exists the possibility that the units interfere with one another, such that one unit's potential outcome when exposed to a specific treatment level, may also depend on the treatment received by another unit. Second, because in multi-unit settings, we must have available more than one copy of each treatment, we may face circumstances in which a unit's potential outcome when receiving the same nominal level of a treatment could vary with different versions of that treatment. These are serious complications, serious in the sense that unless we restrict them by assumptions, combined with careful study design to make these assumptions more realistic, any causal inference will have only limited credibility.

Throughout most of this book, we shall maintain SUTVA. In some cases, however, specific information may suggest that alternative assumptions are more appropriate. For example, in some early AIDS drug trial settings, many patients took some of their assigned drug and shared the remainder with other patients in hopes of avoiding placebos. Given this knowledge, it is clearly no longer appropriate to assert the no-interference element of SUTVA – that treatments assigned to one unit do not affect the outcomes for others. We can, however, use this specific information to model how treatments are received across patients in the study, making alternative – and in this case, more appropriate – assumptions that allow some inference. For example, SUTVA may

be more appropriate using subgroups of people as units in such AIDS drug trials. Similarly, in educational settings, SUTVA may be more plausible with classrooms or schools as the units of analysis than with students as the units of analysis. In many economic examples, interactions between units are often modeled through assumptions on market structure, again avoiding the no-interference element of SUTVA. Consequently, SUTVA is only one candidate exclusion restriction for modeling the potentially complex interactions between units and the entire set of treatment levels in a particular experiment. In many settings, however, it appears that SUTVA is the leading choice.

1.7 THE ASSIGNMENT MECHANISM: AN INTRODUCTION

If we are willing to accept SUTVA, our complicated "You" and "I" aspirin example simplifies to the situation depicted in Table 1.3. Now You and I each face only two treatment levels (e.g., for "You" whether or not "You" take an aspirin), and the accompanying potential outcomes are a function of only our individual actions. This extends readily to many units. To accommodate this generalization, and also the discussion of other examples beyond that of taking or not taking aspirin, as introduced in Section 1.6, let us index the units in the population of size N by i, taking on values $1, \ldots, N$, and let the treatment indicator W_i take on the values 0 (the control treatment, e.g., no aspirin) and 1 (the active treatment, e.g., aspirin). We have one realized (and possibly observed) potential outcome for each unit. For unit i, now $i \in \{1, \ldots, N\}$, let Y_i^{obs} denote this realized (and possibly observed) outcome:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

For each unit we also have one missing potential outcome, for unit i denoted by Y_i^{mis}:

$$Y_i^{\text{mis}} = Y_i(1 - W_i) = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1. \end{cases}$$

Many writers replace the potential outcomes and treatment indicator with simply the treatment indicator, W_i, and the observed outcome Y_i^{obs}. This "observed-value" notation confuses the objects of inference and the assignment mechanism and can lead to mistakes as we see in Section 1.9.

This information alone, still, does not allow us to infer the causal effect of taking an aspirin on headaches. Suppose, in the two-person headache example, that the person who chose not to take the aspirin did so because he had only a minor headache. Suppose then that an hour later both headaches have faded: the headache for the first person possibly faded because of the aspirin (it would still be there without the aspirin), and the headache of the second person faded simply because it was not a serious headache (it would be gone even without the aspirin). When comparing these two observed potential outcomes, we might conclude that the aspirin had no effect, whereas in fact it may have been the cause of easing the more serious headache. The key piece of information that

Table 1.3. ***Example of Potential Outcomes and Causal Effects under SUTVA with Two Units***

Unit	Unknown			Known	
	Potential Outcomes		Causal Effect	Actual	Observed
	Y(Aspirin)	Y(No Aspirin)		Treatment W_i	Outcome Y_i^{obs}
You	No Headache	Headache	Improvement due to Aspirin	Aspirin	No Headache
I	No Headache	No Headache	None	No Aspirin	No Headache

Table 1.4. ***Medical Example with Two Treatments, Four Units, and SUTVA: Surgery (S) and Drug Treatment (D)***

Unit	Potential Outcomes		Causal Effect
	$Y_i(0)$	$Y_i(1)$	$Y_i(1) - Y_i(0)$
Patient #1	1	7	6
Patient #2	6	5	−1
Patient #3	1	5	4
Patient #4	8	7	−1
Average	4	6	2

we lack is how each individual came to receive the treatment level actually received: in our language of causation, the *assignment mechanism*.

Because causal effects are defined by comparing potential outcomes (only one of which can ever be observed), they are well defined irrespective of the actions actually taken. But, because we observe at most half of all potential outcomes, and none of the unit-level causal effects, there is an inferential problem associated with assessing causal effects. In this sense, the problem of causal inference is, as pointed out in Rubin (1974), a *missing data problem*: given any treatment assigned to an individual unit, the potential outcome associated with any alternate treatment is missing. A key role is therefore played by the missing data mechanism, or, as we refer to it in the causal inference context, the assignment mechanism. How is it determined which units get which treatments or, equivalently, which potential outcomes are realized and which are not? This mechanism is, in fact, so crucial to the problem of causal inference that Parts II through VI of this book are organized by varying assumptions concerning this mechanism.

To illustrate the critical role of the assignment mechanism, consider the simple hypothetical example in Table 1.4. This example involves four units, in this case patients, and two possible medical procedures labeled 0 (Drug) and 1 (Surgery). Assuming SUTVA, Table 1.4 displays each patient's potential outcomes, in terms of years of post-treatment survival, under each treatment. From Table 1.4, it is clear that on average, Surgery is better than Drug by two years' life expectancy, that is, the average causal effect of Surgery versus Drug is two years for these four individuals.

Suppose now that the doctor, through expertise or magic, knows enough about these potential outcomes and so assigns each patient to the treatment that is more beneficial to that patient. In this scenario, Patients 1 and 3 will receive surgery, and Patients 2 and

Table 1.5. ***Ideal Medical Practice: Patients Assigned to the Individually Optimal Treatment; Example from Table 1.4***

Unit i	Treatment W_i	Observed Outcome Y_i^{obs}
Patient #1	1	7
Patient #2	0	6
Patient #3	1	5
Patient #4	0	8

4 will receive the drug treatment. The observed treatments and outcomes will then be as displayed in Table 1.5, where the average observed outcome with surgery is one year less than the average observed outcome with the drug treatment. Thus, a casual observer might be led to believe that, on average, the drug treatment is superior to surgery. In fact, the opposite is true: as shown in Table 1.4, if the drug treatment were uniformly applied to a population like these four patients, the average survival would be four years, as can be seen from the "$Y(0)$" column in Table 1.4, as opposed to six years if all patients were treated with surgery, as can be seen from the "$Y(1)$" column in the same table. Based on this example, we can see that we cannot simply look at the observed values of potential outcomes under different treatments, that is, $\{Y_i^{\text{obs}} | i : \text{ s.t. } W_i = 0\}$ and $\{Y_i^{\text{obs}} | i : \text{ s.t. } W_i = 1\}$, and reach valid causal conclusions irrespective of the assignment mechanism. In order to draw valid causal inferences, we must consider why some units received one treatment rather than another. In Parts II through VI of this text, we will discuss in greater detail various assignment mechanisms and the accompanying analyses for drawing valid causal inferences.

1.8 ATTRIBUTES, PRE-TREATMENT VARIABLES, OR COVARIATES

Consider a study of causal effects involving many units, which we assume satisfies the stability assumption, SUTVA. At least half of all potential outcomes will be unobserved or missing, because only one potential outcome can be observed for each unit, namely the potential outcome corresponding to the realized level of the treatment or action. To estimate the causal effect for any particular unit, we will generally need to predict, or impute, the missing potential outcome. Comparing the imputed missing outcome to the realized and observed outcome for this unit allows us to estimate the unit-level causal effect. In general, creating such predictions is difficult. They involve assumptions about the assignment mechanism and about comparisons between different units, each exposed to only one of the treatments. Often the presence of unit-specific background attributes, also referred to as pre-treatment variables, or covariates, and denoted in this text by the K-component row vector X_i for unit i, can assist in making these predictions. For instance, in our headache example, such variables could include the intensity of the headache before making the decision to take aspirin or not. Similarly, in an evaluation of the effect of job training on future earnings, these attributes may include age, previous educational achievement, family, and socio-economic status, or pre–training earnings.

As these examples illustrate, sometimes a covariate (e.g., pre-training earnings) differs from the potential outcome (post-training earnings) solely in the timing of measurement, in which case the covariates can be highly predictive of the potential outcomes.

The key characteristic of these covariates is that they are *a priori* known to be unaffected by the treatment assignment. This knowledge often comes from the fact that they are permanent characteristics of units, or that they took on their values prior to the treatment being assigned, as reflected in the label "pre-treatment" variables.

The information available in these covariates can be used in three ways. First, covariates commonly serve to make estimates more precise by explaining some of the variation in outcomes. For instance, in the headache example, holding constant the intensity of the headache before receiving the treatment by studying units with the same initial headache intensity should give more precise estimates of the effect of aspirin, at least for units with that level of headache intensity. Second, for substantive reasons, the researcher may be interested in the typical (e.g., average) causal effect of the treatment on subgroups (as defined by a covariate) in the population of interest. For example, we may want to evaluate the effects of a job-training program separately for people with different education levels, or the effect of a medical drug separately for women and men. The final and most important role for covariates in our context, however, concerns their effect on the assignment mechanism. Young unemployed individuals may be more interested in training programs aimed at acquiring new skills, or high-risk groups may be more likely to take flu shots. As a result, those taking the active treatment may differ in the values of their background characteristics from those taking the control treatment. At the same time, these characteristics may be associated with the potential outcomes. As a result, assumptions about the assignment mechanism and its possible freedom from dependence on potential outcomes are typically more plausible within subpopulations that are homogeneous with respect to some covariates, that is, conditionally given the covariates, rather than unconditionally.

1.9 POTENTIAL OUTCOMES AND LORD'S PARADOX

To illustrate the clarity that comes with the potential outcomes interpretation of causality, we consider a problem from the literature that is known as Lord's paradox:

> A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded. (Lord, 1967, p. 304)

The results of the hypothetical study described in Lord's paper include the finding that for the males the average weight is identical at the end of the school year to what it was at the beginning; in fact, the whole distribution of weights is unchanged, although some males lost weight and some males gained weight – the gains and losses exactly balance. The same thing is true for the females. The only difference is that the females started and ended the year lighter on average than the males. On average, there is no weight gain or weight loss for either males or females. From Lord's quoted description of the problem, the object of interest, what we will generally call the *estimand*, is the difference between

the causal effect of the university diet on males and the causal effect of the university diet on females. That is, the causal estimand is the difference between the causal effects for males and females, the "differential" causal effect.

The paradox is generated by considering the contradictory conclusions of two statisticians asked to comment on the data. Statistician 1 observes that there are no differences between the September and June weight distributions for either males or females. Thus, Statistician 1 concludes that

> as far as these data are concerned, there is no evidence of any interesting effect of diet (or of anything else) on student weight. In particular, there is no evidence of any differential effect on the two sexes, since neither group shows any systematic change. (Lord, 1967, p. 305)

Statistician 2 looks at the data in a more "sophisticated" way. Effectively, he examines males and females with the same initial weight in September, say a subgroup of "overweight" females (meaning simply above-average-weight females) and a subgroup of "underweight" males (analogously defined). He notices that these males tended to gain weight on average and these females tended to lose weight on average. He also notices that this result is true no matter what the value of initial weight he focuses on. (Actually, Lord's Statistician 2 used a technique known as covariance adjustment or regression adjustment described in Chapter 7.) His conclusion, therefore, is that after "controlling for" initial weight, the diet has a differential positive effect on males relative to females because for males and females with the same initial weight, on average the males gain more than the females.

Who's right? Statistician 1 or Statistician 2? Notice the focus of both statisticians on before-after or gain scores and recall that such gain scores are not causal effects because they do not compare potential outcomes at the same time post-treatment; rather, they compare changes over time. If both statisticians confined their comments to *describing* the data, both would be correct, but for causal inference, both are wrong because these data cannot support any conclusions about the causal effect of the diet without making some very strong, and arguably implausible, assumptions.

Back to the basics. The units are obviously the students, and the time of application of active treatment (the university diet) is clearly September and the time of the recording of the outcome Y is clearly June. Let us accept the stability assumption. Now, what are the potential outcomes, and what is the assignment mechanism? Notice that Lord's statement of the problem uses the already criticized notation with a treatment indicator and the observed variable, Y_i^{obs}, rather than the potential outcome notation being advocated. The potential outcomes are June weight under the university diet $Y_i(1)$ and under the "control" diet $Y_i(0)$. The covariates are sex of students, male versus female, and September weight. But the assignment mechanism has assigned everyone to the new treatment! There is no one, male or female, who is assigned to the control treatment. Hence, there is absolutely no purely empirical basis on which to compare the effects, either raw or differential, of the university diet with the control diet. By making the problem complicated with the introduction of the covariates "male/female" and "initial weight," Lord has created partial confusion. But the point here is that the "paradox" is immediately resolved through the explicit use of potential outcomes. Either answer could be correct for causal inference depending on what we are willing to assume about

the (never-observed) potential outcome under the control diet and its relation to the (observed) potential outcome given the university diet.

1.10 CAUSAL ESTIMANDS

Let us now be a little more formal when describing causal estimands, the ultimate object of interest in our analyses. We start with a population of units, indexed by $i = 1, \ldots, N$, which is our focus. Each unit in this population can be exposed to one of a set of treatments. In the most general case, let $\mathbb{T}_i$ denote the set of treatments to which unit i can be exposed. In most cases, this set will be identical for all units. Exceptions include settings where the treatment is defined as the peer group for each individual. In the current text, the set $\mathbb{T}_i$ consists of the same two treatments for each unit (e.g., taking or not taking a drug),

$$\mathbb{T}_i = \mathbb{T} = \{0, 1\},$$

for all $i = 1, \ldots, N$. Generalizations of most of the discussion in this text to finite sets of treatments are conceptually straightforward.

For each unit i, and for each treatment in the common set of treatments, $\mathbb{T} = \{0, 1\}$, there are corresponding potential outcome, $Y_i(0)$ and $Y_i(1)$. Comparisons of $Y_i(1)$ and $Y_i(0)$ are *unit-level causal effects*. Often these are simple differences,

$$Y_i(1) - Y_i(0), \qquad \text{or ratios } Y_i(1)/Y_i(0),$$

but in general the comparisons can take different forms. There are many such unit-level causal effects, and we often wish to summarize them for the finite sample or for subpopulations. A leading example of what we in general refer to as a *causal estimand* is the average difference of the pair of potential outcomes, averaged over the entire population,

$$\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \big(Y_i(1) - Y_i(0)\big),$$

where the subscript "fs" indicates that we average over the finite sample.

We can generalize this example in a number of ways. Here we discuss two of these generalizations, maintaining in each case the setting with $\mathbb{T} = \{0, 1\}$ for all units. First, we can average over subpopulations rather than over the full population. The subpopulation that we average over may be defined in terms of different sets of variables. First, it can be defined in terms of pre-treatment variables, or covariates, denoted by X_i. Recall these are variables measured on the units that, unlike outcomes, are *a priori* known to be unaffected by the treatment. For example, we may be interested in the average effect of a new drug only for females:

$$\tau_{\text{fs}}(f) = \frac{1}{N(f)} \sum_{i:X_i=f} \big(Y_i(1) - Y_i(0)\big).$$

Here $X_i \in \{f, m\}$ is an indicator for being female, and $N(f) = \sum_{i=1}^{N} \mathbf{1}_{X_i=f}$ is the number of females in the finite population, where $\mathbf{1}_A$ is the indicator function for the event A, equal to 1 if A is true and zero otherwise. Second, one can focus on the average effect of the treatment for those who were exposed to it:

$$\tau_{\text{fs},t} = \frac{1}{N_t} \sum_{i:W_i=1} \big(Y_i(1) - Y_i(0)\big),$$

where N_t is the number of units exposed to the active treatment. For example, we may be interested in the average effect of serving in the military on subsequent earnings in the civilian labor market for those who served in the military, or the average effect of exposure to asbestos on health for those exposed to it. In both examples, there is less interest in the average effect for units not exposed to the treatment. A third way of defining the relevant subpopulation is to do so partly in terms of potential outcomes. As an example, one may be interested in the average effect of a job-training program on hourly wages, averaged only over those individuals who would have been employed (with positive hourly wages) irrespective of the level of the treatment:

$$\tau_{\text{fs,pos}} = \frac{1}{N_{\text{pos}}} \sum_{i:Y_i(0)>0,Y_i(1)>0} \big(Y_i(1) - Y_i(0)\big),$$

where $N_{\text{pos}} = \sum_{i=1}^{N} \mathbf{1}_{Y_i(0)>0,Y_i(1)>0}$. Because the conditioning variable (being employed irrespective of the treatment level) is a function of potential outcomes, the conditioning is (partly) on potential outcomes.

As a second generalization of the average treatment effect, we can focus on more general functions of potential outcomes. For example, we may be interested in the median (over the entire population or over a subpopulation) of $Y_i(1)$ versus the median of $Y_i(0)$. One may also be interested in the median of the difference $Y_i(1) - Y_i(0)$, which generally differs from the difference in medians.

In all cases with $\mathbb{T} = \{0, 1\}$, we can write the causal estimand as a row-exchangeable function of all potential outcomes for all units, all treatment assignments, and pretreatment variables:

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}).$$

In this expression $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are the N-component column vectors of potential outcomes with ith elements equal to $Y_i(0)$ and $Y_i(1)$, $\mathbf{W}$ is the N-component column vector of treatment assignments, with ith element equal to W_i, and $\mathbf{X}$ is the $N \times K$ matrix of covariates with ith row equal to X_i. Not all such functions necessarily have a causal interpretation, but the converse is true: all the causal estimands we consider in this book can be written in this form, and all such estimands are comparisons of $Y_i(0)$ and $Y_i(1)$ for all units in a common set whose definition, as the previous examples illustrate, may depend on $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\mathbf{X}$, and $\mathbf{W}$.

1.11 STRUCTURE OF THE BOOK

The remainder of Part I of this text includes a brief historical overview of the development of our framework for causal inference (Chapter 2) and some mathematical definitions that characterize assignment mechanisms (Chapter 3).

Parts II through V of this text cover different situations corresponding to different assumptions concerning the assignment mechanism. Part II deals with the inferentially simplest setting of randomized assignment, specifically what we call *classical randomized experiments*. In these settings, the assignment mechanism is under the control of the experimenter, and the probability of any assignment of treatments across the units in the experiment is entirely knowable before the experiment begins.

In Parts III and IV we discuss *regular assignment mechanisms*, where the assignment mechanism is not necessarily under the control of the experimenter, and the knowledge of the probabilities of assignment is incomplete in a very specific and limited way: within subpopulations of units defined by fixed values of the covariates, the assignment probabilities are known to be identical for all these units and known to be strictly between zero and one; the probabilities themselves need not be known. Moreover, in practice, we typically have few units with the same values for the covariates, so that the methods discussed in the chapters on classical randomized experiments are not directly applicable.

Finally, Parts V and VI concern *irregular assignment mechanisms*, which allow the assignment to depend on covariates *and* on potential outcomes, both observed and unobserved, or which allow the unit-level assignment probabilities to be equal to zero or one. Such assignment mechanisms present special challenges, and without further assumptions, only limited progress can be made. In this part of the text, we discuss several strategies for addressing these complications in specific settings. For example, we discuss investigating the sensitivity of the inferential results to violations of the critical "unconfoundedness" assumption on the assignment mechanism. We also discuss some specific cases where this unconfoundedness assumption is supplemented by, or replaced by, assumptions linking various potential outcomes. These assumptions are again exclusion restrictions, where specific treatments are assumed *a priori* not to have any, or limited, effects on outcomes. Because of the complications arising from these irregular assignment mechanisms, and the many forms such assignment mechanisms can take in practice, this area remains a fertile field for methodological research.

1.12 SAMPLES, POPULATIONS, AND SUPER-POPULATIONS

In much of the discussion in this text, the finite set of units for which we observe covariates, treatments, and realized outcomes is the set of units we are interested in, and we will refer to this as the *population*. It does not matter how this population was selected, or where it came from. All conclusions are conditional on this population, and we do not attempt to draw inferences for other populations. For part of the discussion, however, it is useful to view the set of units for which we observe values as drawn randomly from a larger population. In that case we typically take the population that the units were drawn from as infinite. When it is important to make this distinction, we will refer to the

set of units for which we observe values as the *finite sample* (often using the subscript "fs"), and the infinite population that these were drawn from as the *super-population* (using subscript "sp") to distinguish between this case and the previous case where we observed values for all units in the population.

1.13 CONCLUSION

In this chapter we present the three basic concepts in our framework for causal inference. The first concept is that of potential outcomes, one for each unit for each level of the treatment. Causal estimands are defined in terms of these potential outcomes, possibly also involving the treatment assignments and pre-treatment variables. We discussed that, because at most only one of the potential outcomes can be observed, there is a need for observing multiple units to be able to conduct causal inference. In order to exploit the presence of multiple units, we use the stability assumption, SUTVA, which is the second basic concept in our framework. The third fundamental concept is that of the assignment mechanism, which determines which units receive which treatment. In Chapter 3 we provide a classification of assignment mechanisms that will serve as the organizing principle of the text.

NOTES

Note that the manipulation underlying our view of causality does not have to take place, merely that one has to be able to do the thought experiment in order for the causal effects to be well defined. Rubin (1978, p. 38) writes: "The fundamental problem facing inference for causal effects is that if treatment t is assigned to the ith experimental unit (i.e., $W_i = t$), only values in Y^t can be observed, Y^j for $j \neq t$ being unobservable (or missing)." Holland (1986, p. 947) puts it similarly when he describes the causal inference problem as arising from the fact that "It is impossible to *observe* the value of $Y_t(u)$ and $Y_c(u)$ on the same unit and, therefore, it is impossible to *observe* the effect of t on u" (emphasis in original). In Holland's notation, u denotes the unit, and $Y_t(u)$ and $Y_c(u)$ denote the two potential outcomes for unit u under the two levels of the treatment. See also Rubin (1977, 2004, 2012).

Following Holland (1986), we refer to the general potential outcomes approach taken in this book as the Rubin Causal Model, although it has precursors in the work by Neyman (1923). Their work explicitly uses potential outcomes ("potential yields" in Neyman, 1990, translation of the 1923 original, p. 467), although Neyman focused exclusively on what we call here completely randomized experiments. In Chapter 2 we discuss in more detail the historical background to the potential outcomes framework.

The Stable Unit Treatment Value Assumption (SUTVA) was formally introduced in Rubin (1980a). See also the discussions in Rubin (1986a, 1990b, 2010). It is implicit in the notation used by Neyman (1923, 1990) where the potential outcomes are indexed only by the treatment assigned to that unit. Cox (1958, p. 19) is explicit about the need for the no-interference part of SUTVA but does not address the part of SUTVA that requires a single version of each treatment for each unit. Fisher does not explicitly address the

issue, but under the null hypothesis of no effect of the treatment whatsoever, SUTVA automatically holds.

For more statistical details of the resolution of Lord's paradox, see Lord (1967) and Holland and Rubin (1983), and for earlier related discussion, see, for example, Lindley and Novick (1981).

There is an extensive econometric literature concerned with causality and methods for inferring causal effects, often in settings with complex selection. For recent reviews, see Angrist and Krueger (2000), Leamer (1988), Heckman and Robb (1984), Heckman, Ichimura, Smith, and Todd, (1998), Heckman, Lalonde, and Smith (2000), and Angrist and Pischke (2008).

Recent textbooks discussing causal inference in various detail and from various points of view include Rosenbaum (1995, 2002, 2009), Shadish, Campbell, and Cook (2002), Van Der Laan and Robins (2003), Lee (2005), Caliendo (2006), Gelman and Hill (2006), Morgan and Winship (2007), Angrist and Pischke (2008), Guo and Fraser (2010), Morton and Williams (2010), Murnane and Willett (2011); and for collected papers, see Rubin (2006) and Freedman (2009). For a more philosophical perspective, see Beebee, Hitchcock, and Menzies (2009). The Rosenbaum books are closest to the current text in terms of the perspective on causality.

There are some approaches to causality that take conceptually different perspectives. In the analysis of time series, economists have found it useful to consider "Granger-Sims causality," which essentially views causality as a prediction property. Suppose we have two time series, one measuring the money supply ("money"), and one measuring gross domestic product (GDP). Money "causes" GDP in the Granger sense if, conditional on the past values of GDP, and possibly conditional on other variables, past values of money predict future values of GDP. Money does not "cause" GDP in the Sims sense if, when predicting money from past, present, and future values of GDP, the future values have no predictive power. See Granger (1969) and Sims (1972). For a recent analysis of the causal links between the money supply (or, more specifically, actions by the Federal Reserve Bank), and GDP, from a perspective that is, at least in spirit, closer to the potential outcome approach taken in this text, see Romer and Romer (2004). Angrist and Kuersteiner (2011) provide some discussion on the link with the potential outcome approach.

Dawid (2000) develops an interesting approach to causality that avoids potential outcomes, and which focuses primarily on a decision-oriented perspective. There has not been much experience with this approach in applications so far.

Pearl (1995, 2000, 2009) advocates a different approach to causality. Pearl combines aspects of structural equations models and path diagrams. In this approach, assumptions underlying causal statements are coded as missing links in the path diagrams. Mathematical methods are then used to infer, from these path diagrams, which causal effects can be inferred from the data, and which cannot. See Pearl (2000, 2009) for details and many examples. Pearl's work is interesting, and many researchers find his arguments that path diagrams are a natural and convenient way to express assumptions about causal structures appealing. In our own work, perhaps influenced by the type of examples arising in social and medical sciences, we have not found this approach to aid drawing of causal inferences, and we do not discuss it further in this text.

CHAPTER 2

A Brief History of the Potential Outcomes Approach to Causal Inference

2.1 INTRODUCTION

The approach to causal inference outlined in the first chapter has important antecedents in the literature. In this chapter we review some of these antecedents to put the potential outcomes approach in perspective. The two most important early developments, in quick succession in the 1920s, are the introduction of potential outcomes in randomized experiments by Neyman (Neyman, 1923, translated and reprinted in Neyman, 1990), and the introduction of randomization as the "reasoned basis" for inference by Fisher (Fisher 1935, p. 14).

Once introduced, the basic idea that causal effects are the comparisons of potential outcomes may seem so obvious that one might expect it to be a long-established tenet of scientific thought. Yet, although the seeds of the idea can be traced back at least to the eighteenth century, the formal notation for potential outcomes was not introduced until 1923 by Neyman. Even then, however, the concept of potential outcomes was used exclusively in the context of randomized experiments, not in observational studies. The same statisticians, analyzing both experimental and observational data with the goal of inferring causal effects, would regularly use the notation of potential outcomes in experimental studies but switch to a notation purely in terms of realized and observed outcomes for observational studies. It is only more recently, starting in the early seventies with the work of Donald Rubin (1974), that the language and reasoning of potential outcomes was put front and center in observational study settings, and it took another quarter century before it found widespread acceptance as a natural way to define and assess causal effects, irrespective of the setting.

Moreover, before the twentieth century there appears to have been only limited awareness of the concept of the assignment mechanism. Although by the 1930s randomized experiments were firmly established in some areas of scientific investigation, notably in agricultural experiments, there was no formal statement for a general assignment mechanism and, moreover, not even formal arguments in favor of randomization until Fisher (1925).

2.2 POTENTIAL OUTCOMES AND THE ASSIGNMENT MECHANISM BEFORE NEYMAN

Before the twentieth century we can find seeds of the potential outcomes definition of causal effects among both experimenters and philosophers. For example, one can see some idea of potential outcomes, although as yet unlabeled as such, in discussions by the philosopher and economist Mill (1973, p. 327), who offers:

> If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten of it, people would be apt to say that eating of that dish was the source of his death.

Applying the potential outcomes notation to this quotation, Mill appears to be considering the two potential outcomes, Y(eat dish) and Y(not eat dish) for the same person. In this case the observed outcome, Y(eat dish), is "death," and Mill appears to posit that if the alternative potential outcome, Y(not eat dish), is "not death," then one could infer that eating the dish was the source (cause) of the death.

Similarly, in the early twentieth century, the father of much of modern statistics, Fisher (1918, p. 214), argued:

> If we say, "This boy has grown tall because he has been well fed," . . . we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.

Here again we see a, somewhat implicit, reference to two potential outcomes, Y(well fed) = tall and Y(not well fed) = shorter, associated with a single unit, a boy.

Despite the insights we may perceive in these quotations, their authors may or may not have intended their words to mean as we choose to interpret them. For instance, in his argument, Mill goes on to require "constant conjunction" in order to assign causality – that is, for the dish to be the cause of death, this outcome must occur every time it is consumed, by this person, or perhaps by any person. Curiously, an early tobacco industry argument used a similar notion of causality: not everyone who smokes two or more packs of cigarettes a day gets lung cancer, therefore smoking does not cause lung cancer. Jerome Cornfield, the well-known American epidemiologist who studied smoking and lung cancer also struggled with this: "If cigarettes are carcinogenic, why don't all smokers get lung cancer?" (Cornfield, 1959, p. 242) without the benefits of the potential outcomes framework. See also Rubin (2012).

No matter how interpreted, however, we have found no early writer who formally pursued these intuitive insights about potential outcomes defining causal effects; in particular, until Neyman did so in 1923, no one developed a formal notation for the idea of potential outcomes. Nor did anyone discuss the importance of the assignment mechanism, which is necessary for the evaluation of causal effects. The first such formal mathematical use of the idea of potential outcomes was introduced by Jerzey Neyman (1923), and then only in the context of an urn model for assigning treatments to plots. The general formal definition of causal effects in terms of potential outcomes, as well as the formal definition of the assignment mechanism, was still another half century away.

2.3 NEYMAN'S (1923) POTENTIAL OUTCOME NOTATION IN RANDOMIZED EXPERIMENTS

Neyman (in the translated 1990 version) begins with a description of a field experiment with m plots on which v varieties might be applied. Neyman introduces what he calls "potential yield" U_{ik}, where i indexes the variety, $i = 1, \ldots, v$, and k indexes the plot, $k = 1, \ldots, m$. The potential yields are not equal to the actual or observed yield because i indexes all varieties and k indexes all plots, and each plot is exposed to only one variety. Throughout, the collection of potential outcomes, $\mathbf{U} = \{U_{ik} : i = 1, \ldots, v;\ k = 1, \ldots, m\}$ is considered *a priori* fixed but unknown. The "best estimate" (Neyman's term) of the yield of the ith variety in the field is the average potential outcomes for that variety over all m plots,

$$a_i = \frac{1}{N} \sum_{k=1}^{m} U_{ik}.$$

Neyman calls a_i the "best estimate" because of his concern with the definition of "true yield," something that he struggled with again in Neyman (1935). As we define potential outcomes, they are the "true" values under SUTVA, not estimates of them.

Neyman then goes on to describe an urn model for determining which variety each plot receives; this model is stochastically identical to the completely randomized experiment with $n = m/v$ plots exposed to each variety. He notes the lack of independence between assignments for different plots implied by this restricted sampling of treatments without replacement (i.e., if plot k receives variety i, then plot l is less likely to receive variety i), and he goes on to note that certain formulas for this situation that have been justified on the basis of independence (i.e., treating the U_{ik} as independent normal random variables given some parameters) need more careful consideration.

Now, still using Neyman's notation, let x_i be the sample average of the n plots actually exposed to the i^{th} variety, as opposed to a_i, the average of the potential outcomes over all m plots. Neyman shows that the expectation of $x_i - x_j$, that is, the average value of $x_i - x_j$ over all assignments that are possible under his urn drawings, is $a_i - a_j$. Thus, the standard estimate of the effect of variety i versus variety j, the difference in observed means, $x_i - x_j$, is unbiased (over repeated randomizations on the m plots) for the causal estimand, $a_i - a_j$, the average effect of variety i versus variety j across all m plots.

Neyman's formalism made three contributions: (*i*) explicit notation for potential outcomes, (*ii*) implicit consideration of something like the stability assumption, and (*iii*) implicit consideration of a model for the assignment of treatments to units that corresponds to the completely randomized experiment. But as Speed (1990, p. 464) writes in his introduction to the translation of Neyman (1923): "Implicit is not explicit; randomization as a physical act, and later as a basis for analysis, was yet to be introduced by Fisher." Nevertheless, the explicit provision of mathematical notation for potential outcomes was a great advance, and after Fisher's introduction of randomized experiments in 1925, Neyman's notation quickly became standard for defining average causal effects in randomized experiments. See, for example, Pitman (1937), Welch (1937), McCarthy (1939), Anscombe (1948), Kempthorne (1952, 1955), Brillinger, Jones, and Tukey (1978), Hedges and Lehman (1970, sec. 9.4), and dozens of other places, often

assuming additivity as in Cox (1956, 1958), and even in introductory texts (Freedman, Pisani, and Purves, 1978, pp. 456–458). Neyman himself, in hindsight, felt that the mathematical model was an advance:

> Neyman has always depreciated the statistical works which he produced in Bydogszcz [which is where Neyman (1923) was done], saying that if there is any merit in them, it is not in the few formulas giving various mathematical expectations but in the construction of a probabilistic model of agricultural trials which, at that time, was a novelty. (Reid, 1982, p. 45)

2.4 EARLIER HINTS FOR PHYSICAL RANDOMIZING

The notion of the central role of randomization, even if not actual randomized experiments, seems to have been "in the air" in the 1920s before it was explicitly introduced by Fisher. For example, "Student" (Gossett, 1923, pp. 281–282) writes: "If now the plots had been randomly placed . . . ," and Fisher and MacKenzie (1923, p. 473) write "Furthermore, if all the plots were undifferentiated, as if the numbers had been mixed up and written down in random order" (see Rubin, 1990, p. 477). Somewhat remarkably, however, an American psychologist and philosopher, Charles Sanders Peirce, appears to have proposed physical randomization decades earlier, although not as a basis for inference, as in Fisher (1925). Specifically, Peirce and Jastrow (1885, reprinted in Stigler, 1980, pp. 75–83) used physical randomization to create sequences of binary treatment conditions (heavier versus lighter weights) in a repeated-measures psychological experiment. The purpose of the randomization was to create sequences such that "any possible psychological guessing of what changes the operator [experimenter] was likely to select was avoided" (Stigler, pp. 79–80).[1] Peirce also appears to have anticipated, in the late nineteenth century, Neyman's concept of unbiased estimation when using simple random samples and appears to have even thought of randomization as a physical process to be implemented in practice (Peirce, 1931).[2] But we can find no suggestion for the physical randomizing of treatments to units as a basis for inference under Fisher (1925).

2.5 FISHER'S (1925) PROPOSAL TO RANDOMIZE TREATMENTS TO UNITS

An interesting aspect of Neyman's analysis was that, as just mentioned, although he developed his notation to treat data as if they arose from what was later called a completely randomly assigned experiment, he did not take the further step of proposing the necessity of physical randomization for credibly assessing causal effects. It was instead Ronald Fisher, in 1925, who first grasped this. Although the distinction may seem trivial in hindsight, Neyman did not see it as such:

---

[1] Thanks to Stephen Stigler for noting this, possibly first, use of randomization in formal experiments, in correspondence with the second author.

[2] Thanks to Keith O'Rourke and Stephen Stigler for pointing this out.

> On one occasion, when someone perceived him as anticipating the English statistician R. A. Fisher in the use of randomization, he objected strenuously:
>
> "I treated *theoretically* an unrestrictedly randomized agricultural experiment and the randomization was considered a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher, and I consider it as one of the most valuable of Fisher's achievements" (Reid, 1982, p. 45)

Also,

> Owing to the work of R. A. Fisher, "Student" and their followers, it is hardly possible to add anything essential to the present knowledge concerning local experiments . . . . One of the most important achievements of the English School is their method of planning field experiments known as the method of Randomized Blocks and Latin Squares. (Neyman, 1935, p. 109)

Thus, independent of Neyman's work, Fisher (1925) proposed the physical randomization of units and furthermore developed a distinct method of inference based for this special class of assignment mechanisms, that is, randomized experiments. The random assignments can be made, for instance, by choosing balls from an urn, as described by Neyman (1923). Fisher's "significance levels" (i.e., p-values), in the current text introduced and discussed in Chapter 5, remain the accepted rigorous standard for the analysis of randomized clinical trials at the start of the twenty-first century and validate so-called *intent-to-treat* analyses, as discussed in Chapters 5 and 23.

2.6 THE OBSERVED OUTCOME NOTATION IN OBSERVATIONAL STUDIES FOR CAUSAL EFFECTS

Despite the almost immediate acceptance of randomized experiments, Fisher's p-values, and Neyman's notation for potential outcomes in agricultural work and mathematical statistics by 1930 within such experiments, these same elements were not used for causal inference in observational studies. Among social scientists, who were using almost exclusively observational data, the work on randomized experiments by Fisher, Neyman, and others, received little or no attention, and researchers continued building models for observed outcomes rather than thinking in terms of potential outcomes. Even among statisticians involved in the analysis of both randomized and non-randomized data for causal effects, the ideas and mathematical language used for causal inference in the setting of randomized experiments were completely excluded from causal inference in the non-randomized settings. The approach in the latter continued to involve building statistical models relating the observed value of the outcome variable to covariates and indicator variables for treatment levels, with the causal effects defined in terms of the parameters of these models, a tradition that appears to originate with Yule (1897).

This approach estimated associations, for example, correlations, between observed variables, and then attempted, using various external arguments about temporal ordering of the variables, to infer causation, that is, to assess which of these associations might be reflecting a causal mechanism. In particular, the pair of the potential outcomes

$(Y_i(1), Y_i(0))$, which in our approach is fundamental for defining causal effects, was replaced by the observed value of Y for unit i, introduced in Section 1.7.

$$Y_i^{\text{obs}} = Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

The observed outcome Y_i^{obs} was then typically regressed, using ordinary least squares methods, as in Yule (1897), on covariates X_i and the indicator for treatment exposure, W_i. The regression coefficient of W_i in this regression was then interpreted as estimating the causal effect of $W_i = 1$ versus $W_i = 0$. Somewhat remarkably, under very specific conditions, this approach works as outlined in Chapter 7. But in broad generality it does not. This tradition dominated economics, sociology, psychology, education, and other social sciences, as well as the biomedical sciences, such as epidemiology, for most of a century.

In fact, for the half century following Neyman (1923), statisticians who wrote with great clarity and insight on randomized experiments using the potential outcomes notation did not use it when discussing non-randomized studies for causal effects. For example, contrast the discussion in Cochran and Cox (1956) on experiments with that in Cochran (1965) on observational studies, and the discussion in Cox (1958) on randomized experiments with that in Cox and McCullagh (1982) on Lord's paradox (which we discussed using the potential outcome framework in Chapter 1).

2.7 EARLY USES OF POTENTIAL OUTCOMES IN OBSERVATIONAL STUDIES IN SOCIAL SCIENCES

Although the potential outcome notation did not find widespread adoption in observational studies until recently, in some specific settings researchers used frameworks for causal inference that are similar. One of the most interesting examples is the use of potential outcomes in the analysis of demand and supply functions specifically, and the analysis of simultaneous equations models in economics in general. In the 1930s and 1940s, economists Tinbergen (1930) and Haavelmo (1944) formulated causal questions in such settings in terms that now appear very modern. Tinbergen writes:

> Let π be any imaginable price; and call total demand at this price $n(\pi)$, and total supply $a(\pi)$. Then the actual price p is determined by the equation $a(p) = n(p)$, so that the actual quantity demanded, or supplied, obeys the condition $u = a(p) = n(p)$, where u is this actual quantity. ... The problem of determining demand and supply curves ... may generally be put as follows: Given p and u as functions of time, what are the functions $n(\pi)$ and $a(\pi)$? (Tinbergen, 1930, translated in Hendry and Morgan, 1994, p. 233)

This quotation clearly describes the potential outcomes and the specific assignment mechanism corresponding to market clearing, closely following the treatment of such questions in economic theory. Note the clear distinction in notation between the price as an argument in the demand-and-supply functon ("any imaginable price π") and the actual price p.

Similarly, Haavelmo (1934) writes:

> If the group of all consumers in society were repeatedly furnished with the total income, or purchasing power r per year, they would, on average or "normally" spend a total amount $\bar{u}$ for consumption per year, equal to $\bar{u} = \alpha r + \beta$. (Haavelmo, 1943, p. 3, reprinted in Hendry and Morgan, 1994, p. 456)

Although more ambiguous than the Tinbergen quote, this certainly suggests that Haavelmo viewed laws or structural equations in terms of potential outcomes that could have been observed by arranging an experiment.

There are two interesting aspects of the Haavelmo work and the link with potential outcomes. First, it appears that Haavelmo was directly influenced by Neyman (see Hendry and Morgan, 1994, p. 67) and in fact studied with him for a couple of months at Berkeley: "I then had the privilege of studying with the world famous statistician Jerzey Neyman for a couple of months in California. ... When I met him for that second talk I had lost most of my illusions regarding my understanding of how to do econometrics" (Haavelmo, 1989). Second, the close connection between the Tinbergen and Haavelmo work and potential outcomes disappeared in later work. In the work by Koopmans and others associated with the Cowles Commission (e.g., the papers in Koopmans, 1950, and Hood and Koopmans, 1953), statistical models are formulated for observed outcomes in terms of observed explanatory variables. No distinction is made between variables that Cox describes as "treatments ... potentially causal" and "intrinsic properties of the [units] under study" (Cox, 1992, p. 296) that are characteristics or attributes of the units. This observed outcome framework for analyzing causal questions dominated economics and other social sciences and continues to dominate the textbooks in econometrics, with few exceptions, until very recently.

2.8 POTENTIAL OUTCOMES AND THE ASSIGNMENT MECHANISM IN OBSERVATIONAL STUDIES: RUBIN (1974)

Rubin (1974, 1975, 1978) makes two key contributions. First, Rubin (1974) puts the potential outcomes center stage in the analysis of causal effects, irrespective of whether the study is an experimental one or an observational one. Second, he discusses the assignment mechanism in terms of the potential outcomes.

Rubin starts by *defining* the causal effect at the unit level in terms of the pair of potential outcomes:

> ... define the causal effect of the E versus C treatment on Y for a particular trial (i.e., a particular unit ...) as follows: Let $y(E)$ be the value of Y measured at t_2 on the unit, given that the unit received the experimental Treatment E initiated at t_1; Let $y(C)$ be the value of Y measured at t_2 on the unit given that the unit recieved the control Treatment C initiated at t_1. Then $y(E) - y(C)$ is the causal effect of the E versus C treatment on Y ... for that particular unit. (Rubin, 1974, p. 639)

This definition fits perfectly with Neyman's framework for analyzing randomized experiments but shows that the definition has nothing to do with the assignment mechanism: it applies equally to observational studies as well as to randomized experiments.

Rubin (1975, 1978) then discusses the benefits of randomization in terms of eliminating systematic differences between treated and control units and formulates the

assignment mechanism in general mathematical terms as possibly depending on the potential outcomes. Our formal consideration of the assignment mechanism begins in Chapter 3.

NOTES

When one of us (Rubin) was visiting the Department of Statistics at Berkeley in the mid-1970s, where Neyman was Professor Emeritus, he asked Neyman why no one ever used the potential outcomes notation from randomized experiments to define causal effects more generally. This meeting was fifteen years before the (re-)publication of Neyman (1923, 1990). Somewhat remarkably in hindsight, at this meeting, Neyman never mentioned that he invented the notation; his reply to the question as to why it was not used outside experiments was to the effect that defining causal effects in non-randomized settings was too speculative, and in such settings, statisticians should stick with statements concerning descriptions and associations (see Rubin, 2010, p. 42). This fits in with the Neyman quote given in Section 2.5: "without randomization, an experiment has little value irrespective of the subsequent treatment" (Reid, 1982, p. 45). The term "assignment mechanism," and its formal definition, including possible dependence on the potential outcomes, was introduced in Rubin (1975).

For discussions on the intention-to-treat principle, see Davies (1954), Fisher et al. (1990), Meier (1992), Cook and DeMets (2008), Wu and Hamada (2009), Altman (1991), Sheiner and Rubin (1995), and Lui (2011).