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." Each of these 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 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 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 job prospects of people with or without college educations, and 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". 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 causal language in statistics. Nevertheless, for many users of statisical methods, causal statements are exactly what they are after.

The fundamental notion underlying our view of causality 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 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, at the same point in time the same unit could have been exposed to an alternative action, treatment, or regime. 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 skin color. Such a pill may not currently exist (but then, neither did surgical procedures for heart transplants two hundred years ago), but we can still contemplate such an action.

This book primarily considers settings with two actions. 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 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 above 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 and 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 now have a different job, and the implied causal statement compares the quality of the job she actually has, to the quality of the job she would have had, 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 ultimately obtained without enrolling in college, is not as well defined as in the first example.

In the third and final 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 as a boy. We could clarify the causal effect by defining such an action in terms of surgical procedures, or hormone treatments, all at various ages, but clearly the causal effect is likely to depend on the specific alternative action 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–formulated as a causal statement.

It may seem restrictive to exclude from consideration such poorly articulated causal questions regarding currently immutable characteristics such as sex or race. However, the reason to do so in our framework is that without further explication of the intervention being con-

sidered, the causal question is not well defined. One can make many of these questions well posed by explicitly articulating the intervention. For example, if the question concerns the causal effect of "race," then an ethnicity change on a *curriculum vitae* (as in Bertrand and Mullainathan, 2004) defines one causal effect being contemplated, whereas if the question concerns a futuristic "at conception change of chromosomees 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.

An alternative 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 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, you, at a particular point in time, contemplating whether or not to take an aspirin for your headache. That is, there are two treatment levels, taking an aspirin, and not taking an aspirin. If you take the aspirin, your 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 $Y(\text{Aspirin})$. (We could use a finer measure of the status of your headache an hour later, for example rating your headache on a ten point scale, but that does not alter the fundamental issues involved here.) Similarly, if you do not take the aspirin, your headache may remain an hour later, or it may not. We denote this potential outcome by $Y(\text{No Aspirin})$, which also can be either "Headache," or "No Headache." There

are therefore two potential outcomes, $Y(\text{Aspirin})$ and $Y(\text{No Aspirin})$, 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:
   $Y(\text{Aspirin}) = \text{No Headache}$, $Y(\text{No Aspirin}) = \text{Headache}$;

2. No effect of aspirin, with a headache in both cases:
   $Y(\text{Aspirin}) = \text{Headache}$, $Y(\text{No Aspirin}) = \text{Headache}$;

3. No effect of aspirin, with the headache gone in both cases:
   $Y(\text{Aspirin}) = \text{No Headache}$, $Y(\text{No Aspirin}) = \text{No Headache}$;

4. Headache gone only without aspirin:
   $Y(\text{Aspirin}) = \text{Headache}$, $Y(\text{No Aspirin}) = \text{No Headache}$.

Table 1.1 illustrates this situation assuming the values $Y(\text{Aspirin}) = \text{No Headache}$, $Y(\text{No Aspirin}) = \text{Headache}$. There is a zero causal effect of taking aspirin in the second and third possibilities.

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 you take an aspirin (and are therefore unable to observe the state of your headache with no aspirin), or do not take an aspirin (and are 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. 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 your headache before and after deciding to take or not take the aspirin. "The fundamental problem facing inference for causal effects" (Rubin, 1978) 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 $Y(\text{Aspirin})$ and will never know the value of $Y(\text{No Aspirin})$ 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 because the woman in question actually went to college, but we will never know what job she would have received if she had decided not to go to college. In general, therefore, even though the unit–level causal effect (here the comparison of your 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 than 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 units 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, why certain potential outcomes where 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 manipulation.

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 he wishes he had 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 illegal, and the causal effect of that action is that a second person suffered damages. From a legal perspective the damages that the second person is entitled to is the difference between 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, which is 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 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 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, you might feel confident that an aspirin is going to relieve your headache within an hour, based on previous experience. Such experiences would include episodes when your headache went away when you took an aspirin, and episodes when your headache did not go away when you did not take aspirin. In that situation, your

views are shaped by comparisons of multiple units: you 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. "You at different times" are not the same unit in our approach to causality. Time matters for many reasons. For example, you may become more or less sensitive to aspirin, evenings may differ from mornings, or the initial intensity of your headache may affect the result. It is often reasonable to assume that time makes little difference for inanimate objects—we may feel confident, based on 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 unit repeatedly, one might observe different physical units 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 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 of 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 most formal studies of causal effects in the social and bio-medical 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 all 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 we both do not. There are therefore four potential outcomes for each of us. For "me" 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 ($6 = \binom{4}{2}$). 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 take an aspirin, or we can compare the status of my headache if only you take an aspirin to the status of my headache 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 and not loose sight of the fact that it is an assumption, 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 mine!

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, 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.) The *Stable Unit Treatment Value Assumption* 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 (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 potenital outcomes.

These two elements of the stability assumption will 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. 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 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 on similar treatments—fundamentally they are assumptions. That is, they rely on previous 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 fertilizer 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 and, with each pair placed in a separate room. In each pair one randomly choosen individual is selected to be the "treated" individual, and the other the

"control" individual. Half the pairs are then randomly selected, with the "treated" individual in the selected pairs given aspirin, and the "treated" individual in the non-selected 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, however, in which the no-interference part of SUTVA is suspect. 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 the same program. 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 assuming that "all 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 immunication 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, e.g., classrooms in educational settings, or specifically limiting the number of units assigned to a particular treatment.

1.6.2 SUTVA: No Hidden Variations in Treatment

The second component of SUTVA, which allows us to exploit the presence of multiple units when assessing causal effects, 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 $Y_i(\text{No Aspirin})$, the weak aspirin outcome $Y_i(\text{Aspirin}-)$ and the strong aspirin outcome $Y_i(\text{Aspirin}+)$. The second part of SUTVA either requires, with treatements labelled "aspirin" and "no-aspirin," or, for example, that the two aspirin outcomes are identical: $Y_i(\text{Aspirin}+) = Y_i(\text{Aspirin}-)$, or that I can only get Aspirin+ and you can only get Aspirin−. Here one way of accounting for this is to define the treatment as taking a randomly selected aspirin (either Aspirin− or Aspirin+). In that case SUTVA might be satisfied for the re-defined stochastic treatment.

Another example of variation in the treatment that is ruled out by SUTVA occurs when differences in the method of administring 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 $w \in \{0, 1\}$. One strategy to make SUTVA more plausible relies on re-defining the represented treatment levels to comprise a larger set of treatments, e.g., 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 as dead or alive rather than for a detailed measurement of health status.

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 assumption, combined with careful study design to make these assumptions more realistic, there are only limited causal inferences that can be drawn.

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 in subgroups of people in such AIDS drug trials. Similarly, 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 modelling 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 make the SUTVA assumption, 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 only of our individual actions. We therefore have one realized (and possibly observed) potential outcome for each unit. For unit i, here "You" or "I", let Y_i^{obs} denote this realized (and possibly observed outcome:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(\text{No Aspirin}) & \text{if } W_i = \text{No Aspirin}, \\ Y_i(\text{Aspirin}) & \text{if } W_i = \text{Aspirin}. \end{cases}$$

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

$$Y_i^{\text{mis}} = \begin{cases} Y_i(\text{Aspirin}) & \text{if } W_i = \text{No Aspirin}, \\ Y_i(\text{No Aspirin}) & \text{if } W_i = \text{Aspirin}. \end{cases}$$

This formulation can be extended without conceptual complications to include many units taking and not taking aspirin.

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 only had 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). In comparing these two observed potential outcomes, we could 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 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 a *missing data problem*: given any treatment assigned to an individual unit, the 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 IV 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 5. This example involves four units, in this case patients, and two possible medical procedures labeled S (Surgery) and D (Drug). Assuming SUTVA, Table 1.4 displays each patient's potential outcomes, in terms of years of post-treatment survival, for 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 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 opposed, as can be seen from the "$Y(D)$" column in Table 1.4, to six years if all patients were treated with surgery, as can be seen from the "$Y(S)$" 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 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 IV of this text, we will discuss in greater detail various assignment mechanisms, and the accompanying analyses for drawing valid causal inferences.

1.8 Attributes, Pretreatment 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 ber realized 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 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 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, can assist in making these predictions. For example, in our headache experiment, such variables could include the intensity of the headache before the treatment was applied. 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, commonly covariates serve to make estimates more precise by controlling for 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 background levels of education, 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 distribution 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 homogenous with respect to some covariates, i.e., conditionally given the covariates, than unconditionally.

1.9 Potential Outcomes and Lord's Paradox

To illustrate the clarity from the potential outcome interpretation of causality, we consider a problem from the literature that is known as Lord's paradox (Lord, 1962):

> "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 arrival in September and the following June are recorded."(Lord, 1962, p.)

The result of the study for the males is that their 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 description of the problem quoted above, the quantity to be estimated, 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. (p. 305)."(Lord, 1962, p.)

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 here 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 gain scores and recall that gain scores are not causal effects because they do not compare potential outcomes at the same time, 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 causal conclusions about the effect of the diet without making some very strong assumptions.

Back to the basics. The units are obviously the students, and the time of application of treatment (the university diet) is clearly September and the time of the recording of the outcome Y is clearly June; accept the stability assumption. Now, what are the potential outcomes and what is the assignment mechanism? Notice that Lord's statement of the problem has reverted to the already criticized observed variable notation, Y^{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 vs. 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 control diet.

1.10 Causal Estimands

Let us now be a little more formal in 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 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. In most of the current text, the set $\mathbb{T}_i$ consists of two treatments for each unit, (e.g., taking or not taking a drug),

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

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 t in their set of treatments, $\mathbb{T}_i$, there is a corresponding potential outcome, $Y_i(t)$. Comparisons of $Y_i(t)$ and $Y_i(s)$, for $s, t \in \mathbb{T}_i$, are unit-level causal effects. Often these are simple differences,

$$Y_i(t) - Y_i(s), \qquad \text{or ratios } Y_i(t)/Y_i(s),$$

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 entire population or for subpopulations. A leading example of what we in general refer to as a *causal estimand* is the average difference of a pair of potential outcomes corresponding to a pair of treatments that

is included in the set of possible treatments for all units, averaged over the entire population,

$$\tau_{\text{avg}}(s,t) = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(t) - Y_i(s)\Big).$$

Such average effects are not always well-defined: in our roommate example there is no treatment t is common to all $\mathbb{T}_i$. In the leading case where $\mathbb{T}_i$ is common to all units, and with only two levels of the treatment, $\mathbb{T}_i = \{0,1\}$, this estimand will be denoted by

$$\tau_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(1) - Y_i(0)\Big).$$

We can generalize this example in a number of ways. Here we discuss two of these generalizations, maintaining in each case the setting with $\mathbb{T}_i = \{0,1\}$ 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 pretreatment variables, or covariates, that is variables measured on the units that are themselves unaffected by the treatment. For example, we may be interested in the average effect of a new drug only for females:

$$\tau_{\text{avg},f} = \frac{1}{N_f}\sum_{i:X_i=f}\Big(Y_i(1) - Y_i(0)\Big),$$

where N_f is the number of females in the population. Second, one can focus on the average effect of the treatment for those who were exposed to it:

$$\tau_{\text{avg,treated}} = \frac{1}{N_t}\sum_{i:W_i=1}\Big(Y_i(1) - Y_i(0)\Big),$$

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. We study specific cases of this in the chapters on *instrumental variables* and *principal stratification* (Chapters 25-27). As an example, one may be interested in the average effect of a job training program on

earnings, averaged only over those individuals who would have been employed (with positive earnings) irrespective of the level of the treatment:

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

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 in general is different from the difference in medians.

In all cases with $\mathbb{T}_i = \{0, 1\}$, 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-vector of potential outcomes, $\mathbf{W}$ is the vector of treatment assignments, and $\mathbf{X}$ is the $N \times K$ matrix of covariates. Not all such functions necessarily have a causal interpretation, but the converse is true: all the causal estimands we consider 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 IV 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 exper-*

iments. In these settings, the assignment mechanism is under the control of the researcher, and the probability of any assignment of treatments across the units in the experiment is entirely known before the experiment begins.

In Part III, we discuss *regular assignment mechanisms*, where the assignment mechanism is not necessarily under the control of the resaercher. However, under regular assignment mechanisms, 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 and known to be strictly between zero and one, although the probabilities themselves need not be known. Moreover, in practice, we often have so few units with the same values for the covariates that the methods discussed in the chapters on classical randomized experiments are not directly applicable.

Finally, Part IV concerns a number of *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, we discuss a number of 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 Conclusion

In this chapter we present the three basic concepts in 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 pretreatment variables. We discussed that, because at most one of the potential outcomes is 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. 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

The quote "there exists no causation without manipulation," is from Rubin (1975), page 238. Note that the manipulation 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, only values in Y^t can be observed, Y^j for $j \neq t$ being unobservable (or missing)." Holland (1986) puts is similarly when writing about the "Fundamental Problem of Causal Inference," as the problem 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." (italics in original, Holland, page 947). 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.

Following Holland (1986), we refer to the 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, page 467), although Neyman focused exclusively on randomized experiments. In Chapter 2 we discuss in more detail the historical background to the potential outcomes framework.

Other views of causality are sometimes used. 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 times series, one mesuring the money supply ("money"), and one measuring gross national product (gdp). Money "causes" gdp in the Granger sense if 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 "causes" gdp in the Sims sense if in a regression of money on past, present and future values of gdp the future values have zero coefficients. 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 (2008) provide some discussion of the link with the potential outcome approach.

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 clearly very interesting, and many researchers find his arguments that path diagrams are a natural and convenient way to express assumptions about causal structures convincing. In our own work, perhaps influenced by the type of examples arising in social and medical sciences, we have not found this approach to facilitate the drawing of causal inferences, and we do not discuss it in detail in this text.

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

Other books:

Morgan and Winship

Angrist and Pischke

Guo and Fraser

Lee

Rosenbaum books

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

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

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

Unit	Unknown Potential Outcomes Y(Aspirin)	Y(No Aspirin)	Causal Effect	Known Actual Treatment	Observed Outcome
You	No Headache	Headache	Improv. due to Asp.	Aspirin	No Headache

Table 1.3: Example of Potential Outcomes and Causal Effects under SUTVA

Unit	Unknown: Potential Outcomes Y(Aspirin)	Y(No Aspirin)	Unknown: Causal Effect	Known: Actual Treatment W_i	Known: Observed Outcome Y_i^{obs}
You	No Headache	Headache	Improv. due to Asp.	Aspirin	No Headache
Me	No Headache	No Headache	None	No Aspirin	No Headache

Table 1.4: Medical Example with Two Treatments: Surgery (S) and Drug Treatment (D)

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

Table 1.5: Ideal Medical Practice: Patients Assigned to the Individually Optimal Treatment

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

Chapter 2

A Brief History of the Potential Outcome 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 outcome approach in perspective. The two most important early developments, in quick succession in the 1920's, are the introduction of potential outcomes in randomized experiments by Jerzey Neyman (Neyman, 1923, reprinted in Neyman, 1990), and the introduction of randomization as the basis for "reasoned inference" by Ronald Fisher (Fisher, 1925, p).

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 18th century, the formal notion of potential outcomes was only introduced in 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 observed outcomes for observational studies. It is only more recently, starting in the early seventies with the work of Rubin (1974) that the language and reasoning of potential outcomes was introduced and used in observational study settings,

and it took another quarter century before it became broadly accepted as the way to define and assess causal effects, irrespective of the setting.

Before the twentieth century there appears to have been limited awareness of the concept of the assignment mechanism. Although by the 1930's 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, no formal arguments in favor of randomization until the work by Ronald Fisher (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 famous philosopher John Stuart Mill, who offers (Mill, 1973, page 327):

> "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, $Y(\text{eat dish})$ and $Y(\text{not eat dish})$ for the same person. In this case the observed outcome, $Y(\text{eat dish})$, is "death", and Mill appears to posit that if the alternative potential outcome, $Y(\text{not eat dish})$, is "not death", then one could infer that eating the dish was the source (cause) of the death.

Similarly, in the early 20^{th}th century, the father of much of modern statistics, Ronald Fisher, argued (1918):

> "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, $Y(\text{well fed}) = \text{tall}$ and $Y(\text{not well fed}) = \text{shorter}$, 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 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.

No matter how interpreted, however, we could find 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 what we now call randomized experiments. 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 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 treated as 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 over all m plots,

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

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 implied by this restricted sampling of treatments without replacement (i.e., it is impossible for one plot to receive more than one variety), and he goes on to note that certain formulas for this situation that have been justified on the basis of independence and/or the Gaussian Law (i.e., treating the U_{ik} as independent normal random variables given some parameters) need more careful consideration.

Now letting x_i be the sample average of the n plots actually exposed to the i^{th} variety, and analogously for x_j, Neyman shows that over all assignments that are possible under his urn drawings, the average value of $x_i - x_j$ 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: (1) explicit notation for potential outcomes, (2) implicit consideration of something like the stability assumption, and (3) 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 when introducing 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 (e.g., Pitman (1937), Welch (1937), McCarthy (1939), Anscombe (1948), Kempthorne (1952, 1955), Brillinger, Jones and Tukey (1978), Hedges and Lehman (1970, Section 9.4), and dozens of other places, often assuming additivity as in Cox (1958), and sometimes being used quite informally as in Freedman, Pisani and Purves (1978, pages 456-458). Neyman himself, in hindsight, felt that the mathematical model was an advance:

> "Neyman has always deprecated 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 Physically Randomizing

The notion of the central role of randomization, but not actual randomized experiments, seems to have been "in the air" in the 1920's. For example, "Student" (Gossett, 1923, pp. 281-282) writes:

> "If now the plots had been randomly placed..." and "...we are as accurate as if we had devoted twice the area to plots randomly arranged."

Somewhat remarkably, however, an American psychologist and philosopher, appears to have proposed physical randomization decades earlier, although not as a basis for inference, as in Fisher (1925). Specifically, Peirce and Jastrow (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."[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 active versus control treatments to units as a basis for inference under Fisher (1925).

2.5 Fisher's (1925) proposal to actually 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

[1] My thanks to Stephen Stigler for calling my attention to this, possibly first, use of randomization in formal experiments.

[2] I owe Keith O'Rourke and Steven Stigler the credit for this scholarship.

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:

> "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, page 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, page 109)."

More specifically, independent of Neyman's work, Fisher (1925) proposed physical randomization of units, and furthermore developed a distinct method of inference based on this special class of assignment mechanisms, subsequently called randomized experiments. The random assignments can be made, for instance, by choosing balls from an urn, as described by Neyman. Fisher's "significance levels" (i.e., p-values), 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 generate so-called "intent-to-treat" analyses (see, for example, Sheiner and Rubin, 1995).

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 primarily observational data, the work on randomized experiments by Neyman received little attention, and researchers continued building models for observed outcomes rather than potential outcomes. Even among statisticians involved in the analysis of both randomized and nonrandomized 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 nonrandomized 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 (1899).

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, were replaced by the observed value of Y for unit i,

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

where $\mathbf{Y}^{\rm obs} = (Y_1^{\rm obs}, \ldots, Y_N^{\rm obs})'$ was typically regressed, using ordinary least squares, on covariates X_i and the indicator for treatment exposure, W_i. The regression coefficient of W_i 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, and other social sciences, as well as the biomedical sciences such as epidemiology, for 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 nonrandomized studies for causal effects. For example, contrast Cochran and Cox (1956), on experiments with Cochran (1965), on observational studies, and Cox (1958), on randomized experiments with Cox and McCullagh (1982), and Lord's paradox discussed in Chapter 1.

2.7 Early Uses of Potential Outcomes 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 very 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.

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 pear year, equal to $\bar{u} = \alpha r + \beta$."(Haavelmo, p., 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, p. 67), and in fact studied with him for a couple of months at Berkeley (). Second, the close connection between the Tinbergen and Haavelmo work and potential outcomes disappeared in later work. In the work by Koopmans (e.g., Koopmans, 1950), and in other work by the Cowles Commission, statistical models are formulated for observed outcomes in terms of observed explanatory variables. No distinction is made between variables that Cox (1992) describes as "potential causal," and "intrinsic variables," that are characteristics of the units. This framework for analyzing causal questions took over in economics, and continuous to dominate the textbooks in econometrics with few exceptions.

2.8 Potential Outcomes and the Assignment Mechanism in Observational Studies: Rubin (1974)

Rubin (1974, 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 "

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 (1978) then discuss the benefits of randomization in terms of eliminating systematic differences between treated and control units, and formulates the assignment mechanism in general in terms of the potential outcomes.

Notes

Even Fisher appeared to fall prey to this confusion when using regression, as pointed out by Rubin (2005, pp. 327–329).

When one of us (Rubin) was visiting the Department of Statistics at Berkeley in the mid-1970's, where Neyman was Professor Emeritus, he was asked 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 non-randomized settings was too speculative, and in such settings, statisticians should stick with statements concerning descriptions and associations. Recall 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).

Chapter 3

A Classification of Assignment Mechanisms

3.1 Introduction

As discussed in Chapter 1, the fundamental problem of causal inference is the presence of missing data—for each unit we can observe at most one of the potential outcomes. A key component in a causal analysis is, therefore, the *assignment mechanism*: the process that determines which units receive which treatments, hence which potential outcomes are realized and thus can be observed, and, conversely, which are missing. In this chapter we introduce a taxonomy of assignment mechanisms that will serve as the organizing principle for this text. Formally, the assignment mechanism describes, as a function of all covariates, and of all potential outcomes, the probability of any vector of assignments. We consider three basic restrictions on assignment mechanisms:

1. individualistic assignment: This limits the dependence of a particular unit's assignment probability on the values of covariates and potential outcomes for other units.

2. probabilistic assignment: This requires the assignment mechanism to imply a non-zero probability for each treatment regime, for every unit.

3. unconfounded assignment: This disallows dependence of the assignment mechanism on the potential outcomes.

Following Cochran (1965), we also make a distinction between experiments, where the assignment mechanism is known and controlled by the researcher, and observational studies, where the assignment mechanism is not under the control of the researcher.

We consider three classes of assignment mechanisms, covered in parts II, III and IV of this

book respectively. The first class, studied in Part II, corresponds to what we call *classical randomized experiments*. Here the assignment mechanism satisfies all three restrictions on the assignment process, and, moreover, the researcher knows and controls the functional form of the assignment mechanism. Such designs are well understood, and in such settings causal effects are of often relatively straightforward to estimate.

We refer to the second class of assignment mechanisms, studied in Part III, as *regular assignment mechanisms*. This class comprises assignment mechanisms that also satisfy the three restrictions, but, in contrast to classical randomized experiments, the assignment mechanism need not be under the control of the researcher. When the assignment mechanism is not under the control of the researcher, the restrictions on the assignment mechanism that make it regular are now usually *assumptions*, and they are typically not satisfied by *design*, as they would be in classical randomized experiments. In general, we will not be sure whether these assumptions hold in any specific application, and in later chapters we will discuss methods for assessing their plausibility, as well as investigate the sensitivity to violations of them. In practice, this type of observational study is of great importance, and it has been studied intensively from a theoretical perspective. Many, but not all, of the methods applicable to randomized experiments can be used, but often modifications to the specific methods are important to enhance the credibility of the results. The simple methods that suffice in the context of randomized experiments tend to be more controversial in this case. The concerns these simple methods raise are particularly serious if the covariate distributions under the various treatment regimes are substantially different, *unbalanced* in our terminology. In that case, it can be very important, for the purpose of making credible causal inferences, to have an initial, what we call *design* stage of the analysis. In this design stage the data on covariate values and treatment assignment (but, importantly, not the final outcome data) are analyzed in order to assemble samples with improved balance in covariate distributions, somewhat in parallel with the design stage of randomized experiments.

In part IV of the book, we analyze some irregular assignment mechanisms, which do not satisfy all three limitations introduced in the first section of this chapter. There is no general theory for irregular designs, and the coverage in this text will focus on settings that have generated particular interest in the theoretical and empirical literatures. One class of assignment mechanisms studied in this part allows for dependence on potential outcomes. Such models have arisen in the econometric literature to account for settings

where individuals choose the treatment regime based on expected benefits associated with the two treatment regimes. Although, as a general matter, such optimizing behavior is not inconsistent with regular assignment mechanisms, in some cases it suggests particular irregular assignment mechanisms associated with so-called *instrumental variable* methods. Other irregular assignment mechanisms include cases where the probability of assignment to treatment versus control is equal to zero or one for some, or even all, units, in what is often described as *regression discontinuity designs*.

The rest of this chapter is organized as follows. In the next section we introduce additional notation. In Section 3.3 we define the assignment mechanism, unit-level assignment probabilities, and the propensity score. In Section 3.4 we formally introduce the three general restrictions we consider imposing on assignment mechanisms. We then use those restrictions to define classical randomized experiments in Section 3.5. In Section 3.6 we define regular assignment mechanisms as a special class of observational studies. The next section, Section 3.7, discusses some non-regular assignment mechanisms. Section 3.8 concludes.

3.2 Notation

Formalizing the potential outcomes discussion in Chapter 1, let us consider a population of N units, indexed by $i = 1, \ldots, N$. The i-th unit in this population is characterized by a K–component vector of covariates (or pre-treatment variables, or attributes), X_i, with $\mathbf{X}$ the $N \times K$ matrix of covariates in the population. In social science applications, the elements of X_i may include an individual's age, education, socioeconomic status, labor market history, pre-test scores, sex and marital status. In biomedical applications, the covariates may include measures of an individual's medical history, and family background information. Most important is that covariates are known *a priori* to be unaffected by the assignment of treatment.

For each unit there is also a pair of potential outcomes, $Y_i(0)$ and $Y_i(1)$, denoting its outcomes under the two values of the treatment: $Y_i(0)$ denotes the outcome under the control treatment, and $Y_i(1)$ denotes the outcome under the active treatment. Notice that in using this notation, we tacitly accept the Stable Unit Treatment Value Assumption (SUTVA) that treatment assignments for other units do not affect the outcomes for unit i, and that each treatment defines a unique outcome for each unit. The latter requirement implies that

there is only a single version of the active and control treatments for each unit. Let $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ denote the N–component vectors (or the N–vectors for short) of the potential outcomes. More generally the potential outcomes could themselves be multi-component vectors, in which case $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ would be matrices with N rows. Here we largely focus on the situation where the potential outcomes are scalars, although in most cases extensions to vector-valued outcomes are conceptually straightforward.

Next, the N-vector of treatment assignments is denoted by $\mathbf{W}$, with i-th element $W_i \in \{0,1\}$, with $W_i = 0$ if unit i received the control treatment, and $W_i = 1$ if this unit received the active treatment. Let $N_c = \sum_{i=1}^{N}(1 - W_i)$ and $N_t = \sum_{i=1}^{N} W_i$ be the number of units assigned to the control and active treatment respectively, with $N_c + N_t = N$.

We do not observe both potential outcomes, and so it is useful to define the observed and missing potential outcomes:

$$Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases} \quad \text{and} \quad Y_i^{\text{mis}} = \begin{cases} Y_i(1) & \text{if } W_i = 0, \\ Y_i(0) & \text{if } W_i = 1. \end{cases} \tag{3.1}$$

$\mathbf{Y}^{\text{obs}}$ and $\mathbf{Y}^{\text{mis}}$ are the corresponding N-vectors (or matrices in the case with multiple outcomes). We can also reverse this relation and characterize the potential outcomes in terms of the observed and missing outcomes:

$$Y_i(0) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 1, \\ Y_i^{\text{obs}} & \text{if } W_i = 0, \end{cases} \quad \text{and} \quad Y_i(1) = \begin{cases} Y_i^{\text{mis}} & \text{if } W_i = 0, \\ Y_i^{\text{obs}} & \text{if } W_i = 1. \end{cases} \tag{3.2}$$

3.3 Assignment Probabilities

To introduce the taxonomy of assignment mechanisms used in this text requires some formal mathematical terms. First, we define the assignment mechanism to be the function that assigns probabilities to all 2^N possible N–vectors of assignments $\mathbf{W}$ (each unit can be assigned to treatment or control), given the N–vectors of potential outcomes $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$, and given the $N \times K$ matrix of covariates $\mathbf{X}$:

Definition 1 (Assignment Mechanism)
Given a population of N units, the assignment mechanism is a row-exchangeable function $\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$*, taking on values in* $[0,1]$*, satisfying*

$$\sum_{\mathbf{W}\in\{0,1\}^N} \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1,$$

for all $\mathbf{X}$, $\mathbf{Y}(0)$, *and* $\mathbf{Y}(1)$.

By the assumption that the function $\Pr(\cdot)$ is row exchangeable, we mean that the order in which we list the N units within the vectors or matrices is irrelevant. Note that this probability $\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is *not* the probability of a particular unit receiving the treatment. Instead it reflects a measure across the full population of N units, for instance the probability that a given assignment vector $\mathbf{W}$—first two units treated, third a control, fourth treated, *etc.*—will occur. The definition requires that the probabilities across the full set of 2^N possible assignment vectors $\mathbf{W}$ sum to one. Note also that some assignment vectors $\mathbf{W}$ may have zero probability. For example, if we were to design a study to evaluate a new drug, it is likely that we would want to rule out the possibility that all subjects received the control treatment. We could do so by assigning zero probability to the vector of assignments $\mathbf{W}$ with $W_i = 0$ for all i, or perhaps even assign zero probability to all vectors of assignments other than those with $\sum_{i=1}^{N} W_i = N/2$, for even values of the sample size N.

In addition to the probability of joint assignment for the entire population, we are often interested in the probability of an individual unit being assigned the treatment:

Definition 2 (Unit Assignment Probability)
The unit level assignment probability for unit i is

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{W}:W_i=1} \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)).$$

Here we sum the probabilities across all possible assignment vectors $\mathbf{W}$ for which $W_i = 1$. Out of the set of 2^N different assignment vectors, half (that is 2^{N-1}) have the property that $W_i = 1$. The probability that unit i is assigned to the control treatment is $1 - p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$. Notice that by this definition, the probability that unit i receives the treatment can be a function of not only its own covariates X_i and potential outcomes $Y_i(1)$ and $Y_i(0)$, but also of the covariate values and potential outcomes of other units in the population.

We are also often interested in the average of the unit-level assignment probabilities for subpopulations with a common value of the covariates. We label this quantity the *propensity score*:

Definition 3 (PROPENSITY SCORE)

The propensity score at x is the average unit assignment probability for units with $X_i = x$,

$$e(x) = \frac{1}{N_x} \sum_{i:X_i=x} p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$$

where $N_x = \#\{i = 1, \ldots, N | X_i = x\}$ is the number of units with $X_i = x$. For values x with $N_x = 0$, the propensity score is defined to be zero.

To illustrate these three definitions more concretely, consider four examples, the first three with with two units, and the last one with three units.

EXAMPLE 1

With two units there are four possible values for $\mathbf{W}$,

$$\mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}.$$

If we are conducting a randomized experiment where all treatment assignments have equal probability the assignment mechanism is equal to

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = 1/4, \quad \text{for } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}. \tag{3.3}$$

In this case the unit assignment probability $p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is equal to $1/2$ for both units $i = 1, 2$. In the absence of covariates, the propensity score is simply a constant, here equal to $1/2$. □

EXAMPLE 2

If instead we are conducting a randomized experiment where only those assignments with exactly one treated and one control unit are allowed, the assignment mechanism is

$$\Pr(\mathbf{W}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \begin{cases} 1/2 & \text{if } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\}, \\ 0 & \text{if } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}. \end{cases} \tag{3.4}$$

This does not change the unit level assignment probability, which remains equal to $1/2$ for both units, and so is the propensity score. □

EXAMPLE 3

A third, more complicated, assignment mechanism with two units is the following. The unit

with the most to gain from the active treatment (using a coin toss in the case of a tie) is assigned to the treatment group, and the other to the control group. This leads to

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = \begin{cases} 1 & \text{if } \mathbf{W} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ and } Y_2(1)-Y_2(0) > Y_1(1)-Y_1(0), \\ 1 & \text{if } \mathbf{W} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } Y_2(1)-Y_2(0) < Y_1(1)-Y_1(0), \\ 1/2 & \text{if } \mathbf{W} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } Y_2(1)-Y_2(0) = Y_1(1)-Y_1(0), \\ 0 & \text{if } \mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}. \end{cases} \tag{3.5}$$

In this example the unit-level treatment probabilities $p_i(\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1))$ are equal to zero, one or a half, depending whether the gain for unit i is smaller or larger than for the other unit, or equal. Given that there are no covariates, the propensity score remains a constant, equal to 1/2 in this case. This is the type of assignment mechanism that we often rule out in attempts to infer causal effects. □

Example 4

This is a sequential randomized experiment. It allows for dependence of the assignment mechanism on the potential outcomes, thus violating some of the assumptions we consider later. In this example there are three units, and thus eight possible values for $\mathbf{W}$,

$$\mathbf{W} \in \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Suppose there is a covariate X_i measuring the order in which the units entered the experiment, $X_i \in \{1,2,3\}$. Without loss of generality, let us assume that $X_i = i$. For the first unit, with $X_i = 1$, a fair coin toss determines the treatment. The second unit, with $X_i = 2$, is assigned to the alternative treatment. Let the observed outcomes for the first and second unit be Y_1^{obs} and Y_2^{obs}. The third unit, with $X_i = 3$, is assigned to the treatment that appears best, based on a comparison of observed outcomes by treatment status for the first two units. If both treatments appear equally beneficial, the third unit is assigned to the active treatment. Forexample, if $W_1 = 0$, $W_2 = 1$, and $Y_1^{\text{obs}} > Y_2^{\text{obs}}$, then the third unit gets

assigned to the control group, if $W_1 = 0$, $W_2 = 1$, and $Y_1^{\rm obs} \leq Y_2^{\rm obs}$, the third units gets assigned to the treatment group, and similarly given the alternative assignments for the first two units. Formally:

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1),\mathbf{X}) = \left\{ \begin{array}{ll} 1/2 & \text{if } \mathbf{W} = \left(\begin{array}{c} 0 \\ 1 \\ 0 \end{array}\right), Y_1(0) > Y_2(1), \\ 1/2 & \text{if } \mathbf{W} = \left(\begin{array}{c} 1 \\ 0 \\ 0 \end{array}\right), Y_1(1) \geq Y_2(0), \\ 1/2 & \text{if } \mathbf{W} = \left(\begin{array}{c} 0 \\ 1 \\ 1 \end{array}\right), Y_1(0) \leq Y_2(1), \\ 1/2 & \text{if } \mathbf{W} = \left(\begin{array}{c} 1 \\ 0 \\ 1 \end{array}\right), Y_1(1) < Y_2(0). \end{array} \right\} \tag{3.6}$$

In this case the unit assignment probability is equal to 1/2 for the first two units, and equal to

$$p_3(\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = \left\{ \begin{array}{ll} 0 & \text{if} Y_1(0) > Y_2(1), \text{ and } Y_1(1) < Y_2(0), \\ 1 & \text{if} Y_1(1) \geq Y_2(0), \text{ and } Y_1(0) \leq Y_2(1), \\ 1/2 & \text{otherwise.} \end{array} \right.$$

for the third unit. Because the covariates identify the unit, the propensity score is equal to the unit assignment probabilities. Thus, for $x = 1$ and $x = 2$ the propensity score is equal to 1/2. If $x = 3$, the propensity score is equal to $p_3(\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1))$. □

3.4 Restrictions on the Assignment Mechanism

Before classifying the various types of assignment mechanisms that are the basis of the organization of the book, let us present three general properties that assignment mechanisms may satisfy. These properties restrict the dependence of the unit-level assignment probabilities on values of covariates and potential outcomes for other units, or restrict the range of values of the unit-level assignment probabilities, or restrict the dependence of the assignment mechanism on potential outcomes.

The first property we consider is *individualistic assignment*, which limits the dependence of the treatment assignment for unit i on the outcomes and assignments for other units:

Definition 4 (Individualistic Assignment)
An assignment mechanism $\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1))$ *is individualistic if, for some function* $q(\cdot)$*, with range* $[0,1]$*,*

$$p_i(\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = q(X_i,Y_i(0),Y_i(1)),$$

and

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = c\cdot\prod_{i=1}^{N} q(X_i,Y_i(0),Y_i(1))^{W_i}\left(1-q(X_i,Y_i(0),Y_i(1))\right)^{1-W_i},$$

for $(\mathbf{W},\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) \in \mathbb{A}$*, for some set* $\mathbb{A}$*, and zero elsewhere. (c is the constant that ensures that the probabilities sum up to unity.)*

This restriction is violated in sequential experiments, e.g., Example 4 earlier. We discuss such cases briefly in the notes to this chapter. Given individualistic assignment, the propensity score simplifies:

$$e(x) = \frac{1}{N_x}\sum_{i:X_i=x} q(X_i,Y_i(0),Y_i(1)).$$

Next, we define *probabilistic assignment*, which requires every unit to have positive probability of being assigned to treatment level 0 and to treatment level 1:

Definition 5 (Probabilistic Assignment)
An assignment mechanism $\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1))$ *is probabilistic if, for each possible* $\mathbf{X}$*,* $\mathbf{Y}(0)$ *and* $\mathbf{Y}(1)$*, the probability of assignment to treatment for unit* i *is strictly between zero and one:*

$$0 < p_i(\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) < 1,$$

for all $i = 1,\ldots,N$*.*

Note that this merely requires that every unit has the possibility of being assigned to the active treatment and the possibility of being assigned to the control treatment.

The third property is a restriction on the dependence of the assignment mechanism on potential outcomes:

Definition 6 (Unconfounded Assignment)
An assignment mechanism is unconfounded if it does not depend on the potential outcomes:

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = \Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}'(0),\mathbf{Y}'(1)),$$

for all $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Y}(0)$, $\mathbf{Y}(1)$, $\mathbf{Y}'(0)$, *and* $\mathbf{Y}'(1)$.

If an assignment mechanism is unconfounded, we can drop the two potential outcomes as arguments and write the assignment mechanism as $\Pr(\mathbf{W}|\mathbf{X})$. Note that the assignment mechanism in Examples 3 and 4 do not satisfy this restriction.

The combination of unconfoundedness and individualistic assignment plays a very important role. In that case, the assignment mechanism and the propensity score simplify considerably:

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = c\cdot\prod_{i=1}^{N} q(X_i)^{W_i}\cdot(1-q(X_i))^{1-W_i}. \tag{3.7}$$

and

$$e(x) = q(x).$$

Thus, under unconfoundedness, the propensity score is not just the average assignment probability for units with covariate value $X_i = x$, it can also be interpreted as the unit-level assignment probability for such units.

Given individualistic assignment, the combination of Assumptions 5 and 6, probabilistic assignment and unconfoundedness, is sometimes referred to as *strongly ignorable treatment assignment* (Rosenbaum and Rubin, 1984). More generally, *ignorable treatment assignment* refers to the case where the assignment mechanism can be written in terms of $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Y}^{\text{obs}}$ only, without dependence on $\mathbf{Y}^{\text{mis}}$.

3.5 Randomized Experiments

Part II of this text deals with the inferentially most straightforward class of assignment mechanisms, randomized assignment. Randomized experimental designs have traditionally been viewed as the most credible basis for causal inference, as reflected in the typical reliance of the Food and Drug Administration in the US on such experiments in its approval process for new pharmaceutical treatments.

Definition 7 (Randomized Experiments)
A randomized experiment is an assignment mechanism that (i) is probabilistic, and, (ii) has a known functional form and is controlled by the researcher.

In Part II of this text we will be concerned with a special case—what we call classical randomized experiments—which additionally assume individualistic assignment and unconfoundedness:

Definition 8 (Classical Randomized Experiments)
A classical randomized experiment is a randomized experiment with an assignment mechanism that is
(i) individualistic,
and (ii) unconfounded.

The definition of a classical randomized experiment rules out sequential experiments as in Example 4. In sequential experiments, the assignment for units assigned in a later stage of the experiment often depend on observed outcomes for units assigned earlier in the experiment.

A leading case of a classical randomized experiment is a *completely randomized experiment*, where, *a priori*, the number of treated units, N_t, is fixed. In such a design, N_t units are randomly selected to receive the active treatment from a population of N units, with the remaining $N_c = N - N_t$ assigned to the control group. In this case, each unit has unit assignment probability $q = N_t/N$, and the assignment mechanism equals

$$\Pr(\mathbf{W}|\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)) = \begin{cases} 1 \Big/ \binom{N}{N_t} & \text{if } \sum_{i=1}^N W_i = N_t, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\binom{N}{N_t} = \frac{N!}{N_t! \cdot (N-N_t)!}.$$

Other prominent examples of classical randomized experiments include stratified randomized experiments and paired randomized experiments.

3.6 Observational Studies: Regular Assignment Mechanisms

In Part III of this text we discuss cases where the exact assignment probabilities may be unknown to the researcher, but the researcher still has substantial information concerning the assignment mechanism. For instance, a leading case will be where the researcher knows the set of variables that enter into the assignment mechanism, but does not know the functional form of the dependence. Such information will generally come from subject matter knowledge. In general we refer to designs with unknown assignment mechanisms as *observational studies*:

Definition 9 (Observational Studies)
An assignment mechanism corresponds to an observational study if the functional form of the assignment mechanism is unknown.

The special case of an observational study that is the main focus of Part III of the book is a *regular assignment mechanism*:

Definition 10 (Regular Assignment Mechanisms)
An assignment mechanism is regular if
(i), the assignment mechanism is individualistic,
(ii), the assignment mechanism is probabilistic, and
(iii), the assignment mechanism is unconfounded.

If, in addition, the functional form of a regular assignment mechanism is known, ther assignment mechanism corresponds to a classical randomized experiment, If the functional form is not known, the assignment mechanism corresponds to an observational study with a regular assignment mechanism.

We shall discuss methods for causal inference for regular assignment mechanisms in great detail. Even if in many cases it may appear too strong to assume that an assignment mechanism is regular, we will argue that in practice it is a very important starting point for many analyses. There are two main reasons for this. The first is that in many well-designed observational studies, researchers have attempted to record all the relevant covariates, that is, all the variables that may be associated with both outcomes and assignment to treatment.

If they have been successful in this endeavor, or at least approximately so, regularity of the assignment mechanism may be a reasonable approximation. The second reason is that specific alternatives to regular assignment mechanisms are typically less credible. Under a regular assignment mechanism, it will be sufficient to adjust appropriately for differences between treated and control units' covariate values to draw valid causal inferences. Any alternative methods will involve causal interpretations of comparisons of units with different treatments who also *differ* systematically in their values for covariates. Rarely is there a convincing argument in support of such alternatives. More details of these arguments are presented in Chapter 12.

3.7 Observational Studies: Irregular Assignment Mechanisms

In Part IV of this book, we discuss a number of irregular assignment mechanisms. Although there is no general approach for all irregular assignment mechanisms, there are a number of interesting and tractable cases that have received much interest in the theoretical and applied literatures. We do not discuss all such cases. In particular, we do not discuss sequential experiments, which, according to our definition, correspond to irregular assignment mechanism. One class of irregular assignment mechanisms we do discuss includes noncompliance in randomized experiments, and relies on *instrumental variables* analyses, and more generally, analyses based on *principal stratification*. Often in these designs, the assignment mechanism is assumed to be "latently regular", that is, it would be regular given some additional covariates that are only partially observed. To conduct inference in such settings, it is often useful to invoke additional conditions, in particular exclusion restrictions, which rule out the presence of particular causal effects.

A second case of interest occurs when the probabilistic assignment assumption—that each unit has a non-zero probability of receiving both treatment levels—is violated. Often such violations compromise any attempt to draw causal inferences. For example, if in the evaluation of a new drug, all men are assigned to the treatment group and all women to the control group, it would be difficult to assess credibly the effect of the new drug. However, in one important case, such violations may not as severely compromise our ability to draw causal inferences as it would in general. Suppose that assignment to treatment is determined

by an eligibility rule that requires, for instance, that a particular covariate fall within some subset of possible values. Specifically, suppose that everybody with a test score belows 50 is assigned to the active treatment, and all others are assigned to the control group. For causal inference, one may, in such *regression discontinuity* settings, compare units with close, but distinct, values of the covariate, relying on smoothness of the relationship between potential outcomes and covariates (test scores) around this threshold value.

A third case of irregular assignment mechanisms where there are some credible approaches to inference requires the presence of *pre* and *post* data for both the treatment and control groups. If the *pre* and *post* data are for the same observational units, it may be reasonable to assume unconfoundedness given the *pre* data, but if the *pre* and *post* data are for different units, this is not feasible. The presence, more generally, of multiple control groups of this type, for example some differing from the treatment group by time and some by location, allows for adjustments based on control group differences, in what is generally known in the econometric literature as *difference in differences* methods.

The remainder of this text will provide more detailed discussion of methods of causal inference given each of these types of assignment mechanisms. In the next part of the book, Chapters 4-11, we start with classical randomized experiments.

3.8 Conclusion

In this chapter we laid out the taxonomy of assignment mechanisms that serves as the organizing principle for this text. Using three restrictions on the assignment mechanism, individualistic assignment, probabilistic assignment, and unconfoundedness, we define regular assignment mechanisms, and the special case of classical randomized experiments. In the next part of the book we will study classical randomized experiments, followed in Part III by the study of observational studies with regular assignment mechanisms. In part IV of the text we will analyse some irregular assignment mechanisms.

Notes

Of the restrictions on assignment mechanisms we discuss in the current chapter the first one, individualistic assignment is often made implicitly, and the term is new. The notion of probabilistic assignment is often stated formally, although it is rarely given a formal label. The term unconfoundedness was coined by Rubin (1990). It is sometimes referred to as the *conditional independence assumption* (Lechner, ; Angrist and Pischke, 2008). In the econometrics literature it is also closely related to the notion of *exogeneity* (Manski, Sandefur, 1992), although formal definitions of exogeneity do not coincide with unconfoundedness (see Imbens, 2004, for some discussion). The combination of probabilistic assignment and unconfoundedness is referred to as *Strong Ignorability* or *Strongly Ignorable Treatment Assignment* by Rosenbaum and Rubin (1984). There is a close link between some of the assumptions used in the context of causal inference and the terminology in missing data problems. In the missing data literature strong ignorability is referred to as *Missing at Random* (Rubin, 1976; Little and Rubin, 2002).

Instrumental variables methods originate in the econometrics literature, and go back to the 1930s and 1940s (Wright 192x, Tinbergen 1928, Haavelmo 1943). For a historical perspective see Stock (). For a modern approach see Imbens and Angrist (1994), and Angrist, Imbens and Rubin (1996). For textbook discussions see Wooldridge () and Angrist and Pischke (2008).

The notion of *Principal Stratification* was introduced by Frangakis and Rubin (2002)

Regression discontinuity designs originate in the psychology literature (Thistlewaite and Campbell, 1960). See for a historical overview Cook (2008), and for a recent survey Lee and Lemieux (2010).

Difference in Differences approaches are widely used in the econometric literature. See Angrist and Pischke (2008) for a general discussion and references.