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 is that of potential outcomes, each corresponding to the various levels of a treatment or manipulation, following the dictum “no causation without manipulation.” Each of these outcomes are a priori observable if the unit were to receive the corresponding treatment level, but ex post, 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,” “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 variables. Nevertheless, statistical theory has been relatively silent on questions of causality. Many
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 statistical methods, causal statements are precisely the goal of their analyses.

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 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. The same object or person at a different time is, for our purposes, a different unit. From this perspective, a causal statement presumes that, although a unit was subject to, or exposed to, a particular action or treatment, at the same point in time an alternative action or treatment could have been taken. For instance, when deciding to take an aspirin to relieve your headache, you could also have chosen not to take the aspirin, or you could have chosen to take an alternative medicine. In this framework, articulating with precision the nature of the action could require 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 occasionally 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 were not taken. The causal effect of the action or treatment involves the comparison of these potential outcomes, one realized (and perhaps observed), and the others not realized and therefore not observed. 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 thought is informally stated as “she has long hair because she is a girl.” In some sense the implicit active action 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, but clearly the causal effect is likely to depend on the specific alternative action chosen. 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 considered, the causal question is not well defined. One can make the question well posed by explicitly articulating the intervention. For example, if the question concerns a “ethnicity change on a curriculum vitae,” then there is one causal effect being contemplated, whereas if the question concerns a futuristic “at conception change of skin color,” there is certainly 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 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 brought about the alternative potential outcome.

### 1.3 The Fundamental Problem of Causal Inference: The Case of a Single Unit

Let us consider the case of a single unit, you, at a particular point in time, contemplating whether to take an aspirin for your headache. That is, there are two treatment levels, taking an aspirin or 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 the 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) comparisons:

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 case 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 cases.

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 outcomes at the same moment in time, whereas the time of the application of the treatment must precede that of the outcome. 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, only one of the potential outcomes can be revealed. 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 \( Y(\text{No Aspirin}) \) but cannot know the value of \( Y(\text{Aspirin}) \). 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 outcome. Table 1.2 illustrates this concept for the aspirin example, assuming the action taken was that you took the aspirin.

For estimation 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 set up, 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.

1.4 The Fundamental Problem of Causal Inference: Multiple Units

Because we cannot learn about causal effects from a single observed 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).

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 may have started of with a more or less severe headache. This type of comparison, often involving many different individuals, is widely used, and it is the basis for most formal studies 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 individuals with, and individuals without, college educations. Similarly, evidence of the effect of smoking on lung cancer is largely based on comparisons of smokers and never smokers, rather than on evidence from individuals who switch from being smokers to being former smokers.

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 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 lose sight of the fact that it is an assumption, not a fact, and therefore may be false. Consider a setting where I will have a headache if you do not take an aspirin, whereas I will not have a headache if you do: we are in the same room, and unless you take an aspirin to ease your own headache, your incessant complaining will maintain mine, regardless of whether I take an aspirin or not!

1.5 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. 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 my headache status. The Stable Unit Treatment Value Assumption incorporates both this idea that units do not interfere with one another, and also 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 potential 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 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 two 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 impossible without such assumptions, and thus it is critical to be explicit about their content and their justifications.

1.5.1 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, for centuries researchers have taken care to separate plots using “guard rows,” unfertilized strips of land between fertilized areas. By controlling the leaching of fertilizer across experimental plots, these guard rows make SUTVA more credible; without them we might suspect that fertilizer applied to one plot affects the yields in contiguous plots.

In our headache example, in order to address the non-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 this directly from a separate experiment in which people are paired and placed in separate rooms, with one member given aspirin and the other not given aspirin. The outcome would then be the status of the headache of the second person in each pair. Although such an experiment could shed some light on the plausibility of our 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 non-interference part of SUTVA can be quite 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 occurs with the effects of innoculations against contagious diseases. The causal effect of your inoculation versus no inoculation will surely depend on the innoculations of others: if everybody else is already innoculated with a perfect vaccine, and others can therefore neither get the disease nor transmit it, your inoculation is superfluous. However, if no one else is innoculated, your treatment can be effective relative to no inoculation. 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.

1.5.2 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 another is new and at full strength. In that case, each of us may in fact have three treatments available: no aspirin, ineffective tablet and 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 that the two aspirin outcomes are identical: $Y(\text{aspirin+}) = Y(\text{aspirin−}),$ 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 is satisfied for the re-defined treatment.

Another example of variation in the treatment that is ruled out occurs when differences in the method of administering the treatment matter. The effect of taking a drug for a particular
individual may differ depending on whether the individual was assigned to receive it, or chose to take it. 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, this 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, does not contain any information regarding 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 $w \in \{0, 1\}$. Strategies to make SUTVA more plausible include re-defining the treatment to comprise a larger set of treatments, or coarsening the outcome to make SUTVA more plausible. For an example of the latter, SUTVA may be more plausible if the outcome is defined as dead or alive rather than for a finer measurement of health status.

### 1.5.3 Alternatives to SUTVA

To summarize the previous discussion, assessing the causal effect of a treatment requires observing more than a single unit, because we must have observations on both potential outcomes: 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 receiving the same nominal level of one treatment could in fact receive different versions of that treatment. These are serious complications, serious in the sense that unless we restrict them, by assumption, and through careful study design and measurement 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 instead 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 conditions, 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.

1.6 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 taken. We therefore now have one realized and possibly observed potential outcome for each unit. 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 first possibly because of the aspirin (it would still be there without the aspirin) and the second 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 decided which units get which treatments, or equivalently, which potential outcomes are observed and which are missing? 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.

Suppose now that the doctor, through expertise or magic, knows all of these potential outcomes (or at least their ranks), 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, an outside 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 to six years if all patients were treated with surgery, as can be seen from the “S” column in Table 1.4. 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.

1.7 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 (that is, missing), because each unit can reveal its potential outcomes under only one 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 obtain an estimate of the unit-level causal effect. In general, creating such predictions is difficult. They involve assumptions about the assignment mechanism, and comparisons between different units exposed to only one of the treatments. Often the presence of unit-specific background attributes, pretreatment 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 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 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 often more plausible within subpopulations that are homogenous with respect to some covariates, i.e., conditionally given the covariates, than unconditionally.

1.8 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 \( T_i \) denote the set of treatments to which unit \( i \) can be exposed to. In most cases, this set will be identical for all units. In most of the current text, the set \( T_i \) consists of two treatments for each unit, (e.g., taking or not taking a drug),

\[
T_i = \{0, 1\},
\]

for all \( i = 1, \ldots, N \). Generalizations of most of the discussion in this text to finite sets of treatments are conceptually straightforward. In some cases, the sets of treatments are more complicated. For example, consider a college assigning its incoming class of \( N \) freshman to \( N/2 \) dorm rooms, with two freshmen assigned to each room. One can think of the set of treatments for each freshman the set of possible roommates, so that

\[
T_i = \{1, \ldots, i-1, i+1, \ldots, N\} = \{1, \ldots, N\} / \{i\}.
\]

Since no person can be assigned to be their own roommate, the set of treatments differs for each individual.
For each unit \( i \), and for each treatment \( t \) in their set of treatments, \( \mathcal{T}_i \), there is a corresponding potential outcome, \( Y_i(t) \). Comparisons of \( Y_i(t) \) and \( Y_i(s) \), for \( s, t \in \mathcal{T}_i \), are unit-level causal effects. Often these are simple differences, 
\[
Y_i(t) - Y_i(s),
\]
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 call a \textit{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,
\[
\tau_{\text{avg}}(s, t) = \frac{1}{N} \sum_{i=1}^{N} (Y_i(t) - Y_i(s)),
\]
which is not well-defined in our roommate example because no treatment \( t \) is common to all \( \mathcal{T}_i \). In the leading case where \( \mathcal{T}_i \) is common to all units, with only two levels of the treatment, \( \mathcal{T}_i = \{0, 1\} \), this estimand will be denoted by
\[
\tau_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0)).
\]
We can generalize this example in a number of ways. Here we discuss three of them, maintaining in each case the setting with \( \mathcal{T}_i = \{0, 1\} \) 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 women:
\[
\tau_{\text{avg},f} = \frac{1}{N_f} \sum_{i: X_i = f} (Y_i(1) - Y_i(0)),
\]
where \( N_f \) is the number of women in the population. Second, one can focus on the average effect of the treatment for those who were exposed to it:
\[
\tau_{\text{avg},\text{treated}} = \frac{1}{N_t} \sum_{i: W_i = 1} (Y_i(1) - Y_i(0)),
\]
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 for those who served in the military, or the average effect of exposure to asbestos 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:

\[
\tau_{\text{avg, pos}} = \frac{1}{N_{\text{pos}}} \sum_{i: Y_i(0) > 0, Y_i(1) > 0} \left( Y_i(1) - Y_i(0) \right),
\]

where \( N_{\text{pos}} = \sum_{i=1}^{N} 1_{Y_i(0) > 0, Y_i(1) > 0} \). 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.

Third, and this is more controversial, we may be interested in averages where the averages are over different levels of the treatment for different units. To see why this may be of interest, consider the roommate example. Suppose for each freshman we observe a binary indicator, say lazy versus studious. There is no well defined causal effect of having a studious versus a lazy roommate for freshman \( i \), because the causal effect is likely to depend on the identity of the roommate. However, for freshman \( i \), we could look at the average difference of having a studious versus a lazy roommate, averaged over all possible studious and lazy roommates in the population. Then we can average this over the entire population of freshman and get a well defined average. Note, however that this average is over different treatments for different freshman.

In all cases with \( T_i = \{0, 1\} \), we can write the causal estimand as a function of all potential
outcomes for all units, all treatment assignments, and pretreatment variables:

\[ \tau = \tau(Y(0), Y(1), X, W). \]

In this expression \( Y(0) \) and \( Y(1) \) are the \( N \)-vector of potential outcomes, \( W \) is the vector of treatment assignments, and \( X \) 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 are comparisons of \( Y_i(0) \) and \( Y_i(1) \) for a common set of units, whose definition, as the previous examples illustrate, may depend on \( Y(0), Y(1), X, \) and \( W \).

1.9 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 cover different situations corresponding to different assumptions concerning the assignment mechanism. Part II deals with the inferentially simplest setting of randomized assignment, specifically what we call classical randomized experiments. In these settings, the assignment mechanism is under the control of the researcher, and thus 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 researcher. However, under regular assignment mechanisms, the knowledge of the probabilities of assignment may be 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, yet the probabilities themselves may 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 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 \textit{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.

\section*{1.10 Conclusion}

In this chapter we present the basic concepts and problems 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 multiple units to be able to conduct causal inference. In order to exploit the presence of multiple units, we use the stability assumption, or 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 the assignment mechanism that will serve as the organizing principle of the text.
Notes

The quote “no causality without manipulation,” is from Rubin (1975).
“the fundamental problem of causal inference” rubin, 1978

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: times series A “causes” time series B if, conditional on the past values of B, and possibly conditional on other variables, past values of A predict future values of B.

Expand

Following Holland (1986), we refer to the view espoused in this book as the Rubin Causal Model, although it has precursors in the work by Fisher () and Neyman () in the context of randomized experiments. as discussed in chapter 2

Potential outcomes are implicit, and sometimes explicit, in many notions of causality. (Haavelmo, Roy, Hurwicz? etc)

Other views: discuss Dawid, Suppes, Pearl.

term potential yields in neyman
Table 1.1: **Example of Potential Outcomes and Causal Effect with One Unit**

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Causal Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(\text{Aspirin})$</td>
<td>$Y(\text{No Aspirin})$</td>
</tr>
<tr>
<td>You</td>
<td>No Headache</td>
<td>Headache</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Unit</th>
<th>Unknown Potential Outcomes</th>
<th>Causal Effect</th>
<th>Known Treatment Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Y(\text{Aspirin})$</td>
<td>$Y(\text{No Aspirin})$</td>
<td>Improv. due to Asp.</td>
</tr>
<tr>
<td>You</td>
<td>No Headache</td>
<td>Headache</td>
<td>Aspirin</td>
</tr>
</tbody>
</table>
Table 1.3: Example of Potential Outcomes and Causal Effects under SUTVA

<table>
<thead>
<tr>
<th>Unit</th>
<th>Unknown Potential Outcomes</th>
<th>Causal Effect</th>
<th>Known Actual Treatment</th>
<th>Observed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>You</td>
<td>Y(Aspirin)</td>
<td>No Headache</td>
<td>Headache</td>
<td>Improv. due to Asp.</td>
</tr>
<tr>
<td>Me</td>
<td>Y(No Aspirin)</td>
<td>No Headache</td>
<td>No Headache</td>
<td>None</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Unit</th>
<th>Potential Outcomes</th>
<th>Causal Effect</th>
<th>S-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient #1</td>
<td>7</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Patient #2</td>
<td>5</td>
<td>6</td>
<td>-1</td>
</tr>
<tr>
<td>Patient #3</td>
<td>5</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Patient #4</td>
<td>7</td>
<td>8</td>
<td>-1</td>
</tr>
<tr>
<td>Average Effect</td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
Table 1.5: Ideal Medical Practice: Patients Assigned to the Individually Optimal Treatment

<table>
<thead>
<tr>
<th>Unit</th>
<th>Treatment</th>
<th>Observed Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient #1</td>
<td>S</td>
<td>7</td>
</tr>
<tr>
<td>Patient #2</td>
<td>D</td>
<td>6</td>
</tr>
<tr>
<td>Patient #3</td>
<td>S</td>
<td>5</td>
</tr>
<tr>
<td>Patient #4</td>
<td>D</td>
<td>8</td>
</tr>
</tbody>
</table>