4

Counterfactuals and Their Applications

4.1 Counterfactuals

While driving home last night, I came to a fork in the road, where I had to make a choice: to take the freeway ($X = 1$) or go on a surface street named Sepulveda Boulevard ($X = 0$). I took Sepulveda, only to find out that the traffic was touch and go. As I arrived home, an hour later, I said to myself: "Gee, I should have taken the freeway."

What does it mean to say, "I should have taken the freeway"? Colloquially, it means, "If I had taken the freeway, I would have gotten home earlier." Scientifically, it means that my mental estimate of the expected driving time on the freeway, on that same day, under the identical circumstances, and governed by the same idiosyncratic driving habits that I have, would have been lower than my actual driving time.

This kind of statement—an "if" statement in which the "if" portion is untrue or unrealized—is known as a *counterfactual*. The "if" portion of a counterfactual is called the *hypothetical condition*, or more often, the *antecedent*. We use counterfactuals to emphasize our wish to compare two outcomes (e.g., driving times) under the exact same conditions, differing only in one aspect: the antecedent, which in our case stands for "taking the freeway" as opposed to the surface street. The fact that we know the outcome of our actual decision is important, because my estimated driving time on the freeway after seeing the consequences of my actual decision (to take Sepulveda) may be totally different from my estimate prior to seeing the consequence. The consequence (1 hour) may provide valuable evidence for the assessment, for example, that the traffic was particularly heavy on that day, and that it might have been due to a brush fire. My statement "I should have taken the freeway" conveys the judgment that whatever mechanisms impeded my speed on Sepulveda would not have affected the speed on the freeway to the same extent. My retrospective estimate is that a freeway drive would have taken less than 1 hour, and this estimate is clearly different than my prospective estimate was, when I made the decision prior to seeing the consequences—otherwise, I would have taken the freeway to begin with.

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

If we try to express this estimate using *do*-expressions, we come to an impasse. Writing

$$E(\textit{driving time}|do(\textit{freeway}), \textit{driving time} = 1\ \textit{hour})$$

leads to a clash between the driving time we wish to estimate and the actual driving time observed. Clearly, to avoid this clash, we must distinguish symbolically between the following two variables:

1. Actual driving time
2. Hypothetical driving time under freeway conditions when actual surface driving time is known to be 1 hour.

Unfortunately, the *do*-operator is too crude to make this distinction. While the *do*-operator allows us to distinguish between two probabilities, $P(\textit{driving time}|do(\textit{freeway}))$ and $P(\textit{driving time}|do(\textit{Sepulveda}))$, it does not offer us the means of distinguishing between the two variables themselves, one standing for the time on Sepulveda, the other for the hypothetical time on the freeway. We need this distinction in order to let the actual driving time (on Sepulveda) inform our assessment of the hypothetical driving time.

Fortunately, making this distinction is easy; we simply use different subscripts to label the two outcomes. We denote the freeway driving time by $Y_{X=1}$ (or Y_1, where context permits) and Sepulveda driving time by $Y_{X=0}$ (or Y_0). In our case, since Y_0 is the Y actually observed, the quantity we wish to estimate is

$$E(Y_{X=1}|X = 0, Y = Y_0 = 1) \tag{4.1}$$

The novice student may feel somewhat uncomfortable at the sight of the last expression, which contains an eclectic mixture of three variables: one hypothetical and two observed, with the hypothetical variable $Y_{X=1}$ predicated upon one event ($X = 1$) and conditioned upon the conflicting event, $X = 0$, which was actually observed. We have not encountered such a clash before. When we used the *do*-operator to predict the effect of interventions, we wrote expressions such as

$$E[Y|do(X = x)] \tag{4.2}$$

The Y in this expression is predicated upon the event $X = x$. With our new notation, the expression might as well have been written $E[Y_{X=x}]$. But since all variables in this expression were measured in the same world, there is no need to abandon the *do*-operator and invoke counterfactual notation.

We run into problems with counterfactual expressions like (4.1) because $Y_{X=1} = y$ and $X = 0$ are—and must be—events occurring under different conditions, sometimes referred to as "different worlds." This problem does not occur in intervention expressions, because Eq. (4.1) seeks to estimate our total drive time in a world where we chose the freeway, given that the actual drive time (in the world where we chose Sepulveda) was 1 hour, whereas Eq. (4.2) seeks to estimate the expected drive time in a world where we chose the freeway, with no reference whatsoever to another world.

In Eq. (4.1), however, the clash prevents us from reducing the expression to a *do*-expression, which means that it cannot be estimated from interventional experiments. Indeed, a randomized controlled experiment on the two decision options will never get us the estimate we want. Such experiments can give us $E[Y_1] = E[Y|do(freeway)]$ and $E[Y_0] = E[Y|do(Sepulveda)]$, but the fact that we cannot take both the freeway and Sepulveda simultaneously prohibits us from estimating the quantity we wish to estimate, that is, the conditional expectation $E[Y_1|X = 0, Y = 1]$. One might be tempted to circumvent this difficulty by measuring the freeway time at a later time, or of another driver, but then conditions may change with time, and the other driver may have different driving habits than I. In either case, the driving time we would be measuring under such surrogates will only be an approximation of the one we set out to estimate, Y_1, and the degree of approximation would vary with the assumptions we can make on how similar those surrogate conditions are to my own driving time had I taken the freeway. Such approximations may be appropriate for estimating the target quantity under some circumstances, but they are not appropriate for *defining* it. Definitions should accurately capture what we wish to estimate, and for this reason, we must resort to a subscript notation, Y_1, with the understanding that Y_1 is my "would-be" driving time, had I chosen the freeway at that very juncture of history.

Readers will be pleased to know that their discomfort with the clashing nature of Eq. (4.1) will be short-lived. Despite the hypothetical nature of the counterfactual Y_1, the structural causal models that we have studied in Part Two of the book will prove capable not only of computing probabilities of counterfactuals for any fully specified model, but also of estimating those probabilities from data, when the underlying functions are not specified or when some of the variables are unmeasured.

In the next section, we detail the methods for computing and estimating properties of counterfactuals. Once we have done that, we'll use those methods to solve all sorts of complex, seemingly intractable problems. We'll use counterfactuals to determine the efficacy of a job training program by figuring out how many enrollees would have gotten jobs had they not enrolled; to predict the effect of an additive intervention (adding 5 mg/l of insulin to a group of patients with varying insulin levels) from experimental studies that exercised a uniform intervention (setting the group of patients' insulin levels to the same constant value); to ascertain the likelihood that an individual cancer patient would have had a different outcome, had she chosen a different treatment; to prove, with a sufficient probability, whether a company was discriminating when they passed over a job applicant; and to suss out, via analysis of direct and indirect effects, the efficacy of gender-blind hiring practices on rectifying gender disparities in the workforce.

All this and more, we can do with counterfactuals. But first, we have to learn how to define them, how to compute them, and how to use them in practice.

4.2 Defining and Computing Counterfactuals

4.2.1 The Structural Interpretation of Counterfactuals

We saw in the subsection on interventions that structural causal models can be used to predict the effect of actions and policies that have never been implemented before. The action of setting a variable, X, to value x is simulated by replacing the structural equation for X with the equation $X = x$. In this section, we show that by using the same operation in a slightly different context,

we can use SEMs to define what counterfactuals stand for, how to read counterfactuals from a given model, and how probabilities of counterfactuals can be estimated when portions of the models are unknown.

We begin with a fully specified model M, for which we know both the functions $\{F\}$ and the values of all exogenous variables. In such a deterministic model, every assignment $U = u$ to the exogenous variables corresponds to a single member of, or "unit" in a population, or to a "situation" in nature. The reason for this correspondence is as follows: Each assignment $U = u$ uniquely determines the values of all variables in V. Analogously, the characteristics of each individual "unit" in a population have unique values, depending on that individual's identity. If the population is "people," these characteristics include salary, address, education, propensity to engage in musical activity, and all other properties we associate with that individual at any given time. If the population is "agricultural lots," these characteristics include soil content, surrounding climate, and local wildlife, among others. There are so many of these defining properties that they cannot all possibly be included in the model, but taken all together, they uniquely distinguish each individual and determine the values of the variables we do include in the model. It is in this sense that every assignment $U = u$ corresponds to a single member or "unit" in a population, or to a "situation" in nature.

For example, if $U = u$ stands for the defining characteristics of an individual named Joe, and X stands for a variable named "salary," then $X(u)$ stands for Joe's salary. If $U = u$ stands for the identity of an agricultural lot and Y stands for the yield measured in a given season, then $Y(u)$, stands for the yield produced by lot $U = u$ in that season.

Consider now the counterfactual sentence, "Y would be y had X been x, in situation $U = u$," denoted $Y_x(u)= y$, where Y and X are any two variables in V. The key to interpreting such a sentence is to treat the phrase "had X been x" as an instruction to make a minimal modification in the current model so as to establish the antecedent condition $X = x$, which is likely to conflict with the observed value of $X, X(u)$. Such a minimal modification amounts to replacing the equation for X with a constant x, which may be thought of as an external intervention $do(X = x)$, not necessarily by a human experimenter. This replacement permits the constant x to differ from the actual value of X (namely, $X(u)$) without rendering the system of equations inconsistent, and in this way, it allows all variables, exogenous as well as endogenous, to serve as antecedents to other variables.

We demonstrate this definition on a simple causal model consisting of just three variables, X, Y, U, and defined by two equations:

$$X = aU \tag{4.3}$$

$$Y = bX + U \tag{4.4}$$

We first compute the counterfactual $Y_x(u)$, that is, what Y would be had X been x, in situation $U = u$. Replacing the first equation with $X = x$ gives the "modified" model M_x:

$$X = x$$

$$Y = bX + U$$

Substituting $U = u$ and solving for Y gives

$$Y_x(u) = bx + u$$

Table 4.1 The values attained by $X(u)$, $Y(u)$, $Y_x(u)$, and $X_y(u)$ in the linear model of Eqs. (4.3) and (4.4)

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

which is expected, since the meaning of the structural equation $Y = bX + U$ is, exactly "the value that Nature assigns to Y must be U plus b times the value assigned to X." To demonstrate a less obvious result, let us examine the counterfactual $X_y(u)$, that is, what X would be had Y been y in situation $U = u$. Here, we replace the second equation by the constant $Y = y$ and, solving for X, we get $X_y(u) = au$, which means that X remains unaltered by the hypothetical condition "had Y been y." This should be expected, if we interpret this hypothetical condition as emanating from an external, albeit unspecified, intervention. It is less expected if we do not invoke the intervention metaphor but merely treat $Y = y$ as a spontaneous, unanticipated change. The invariance of X under such a counterfactual condition reflects the intuition that hypothesizing future eventualities does not alter the past.

Each SCM encodes within it many such counterfactuals, corresponding to the various values that its variables can take. To illustrate additional counterfactuals generated by this model, let us assume that U can take on three values, 1, 2, and 3, and let $a = b = 1$ in Eqs. (4.3) and (4.4). Table 4.1 gives the values of $X(u)$, $Y(u)$, $Y_x(u)$, and $X_y(u)$ for several levels of x and y. For example, to compute $Y_2(u)$ for $u = 2$, we simply solve a new set of equations, with $X = 2$ replacing $X = aU$, and obtain $Y_2(u) = 2 + u = 4$. The computation is extremely simple, which goes to show that, while counterfactuals are considered hypothetical, or even mystical from a statistical view point, they emerge quite naturally from our perception of reality, as encoded in structural models. Every structural equation model assigns a definitive value to every conceivable counterfactual.

From this example, the reader may get the impression that counterfactuals are no different than ordinary interventions, captured by the *do*-operator. Note, however, that, in this example we computed not merely the probability or expected value of Y under one intervention or another, but the actual value of Y under the hypothesized new condition $X = x$. For each situation $U = u$, we obtained a definite number, $Y_x(u)$, which stands for that hypothetical value of Y in that situation. The *do*-operator, on the other hand, is only defined on probability distributions and, after deleting the factor $P(x_i|pa_i)$ from the product decomposition (Eq. (1.29)), always delivers probabilistic results such as $E[Y|do(x)]$. From an experimentalist perspective, this difference reflects a profound gap between population and individual levels of analysis; the $do(x)$-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual, $U = u$, under such interventions. This difference has far-reaching consequences, and will enable us to define probabilities of concepts such as credit, blame, and regret, which the *do*-operator is not able to capture.

4.2.2 The Fundamental Law of Counterfactuals

We are now ready to generalize the concept of counterfactuals to any structural model, M. Consider any arbitrary two variables X and Y, not necessarily connected by a single equation.

Let M_x stand for the modified version of M, with the equation of X replaced by $X = x$. The formal definition of the counterfactual $Y_x(u)$ reads

$$Y_x(u) = Y_{M_x}(u) \tag{4.5}$$

In words: The counterfactual $Y_x(u)$ in model M is defined as the solution for Y in the "surgically modified" submodel M_x. Equation (4.5) is one of the most fundamental principles of causal inference. It allows us to take our scientific conception of reality, M, and use it to generate answers to an enormous number of hypothetical questions of the type "What would Y be had X been x?" The same definition is applicable when X and Y are sets of variables, if by M_x we mean a model where the equations of all members of X are replaced by constants. This raises enormously the number of counterfactual sentences computable by a given model and brings up an interesting question: How can a simple model, consisting of just a few equations, assign values to so many counterfactuals? The answer is that the values that these counterfactuals receive are not totally arbitrary, but must cohere with each other to be consistent with an underlying model.

For example, if we observe $X(u) = 1$ and $Y(u) = 0$, then $Y_{X=1}(u)$ must be zero, because setting X to a value it already has, $X(u)$, should produce no change in the world. Hence, Y should stay at its current value of $Y(u) = 0$.

In general, counterfactuals obey the following *consistency rule*:

$$\textit{if} \quad X = x \quad \textit{then} \quad Y_x = Y \tag{4.6}$$

If X is binary, then the consistency rule takes the convenient form:

$$Y = XY_1 + (1 - X)Y_0$$

which can be interpreted as follows: Y_1 is equal to the observed value of Y whenever X takes the value one. Symmetrically, Y_0 is equal to the observed value of Y whenever X is zero. All these constraints are automatically satisfied if we compute counterfactuals through Eq. (4.5).

4.2.3 *From Population Data to Individual Behavior—An Illustration*

To illustrate the use of counterfactuals in reasoning about the behavior of an individual unit, we refer to the model depicted in Figure 4.1, which represents an "encouragement design": X represents the amount of time a student spends in an after-school remedial program, H the amount of homework a student does, and Y a student's score on the exam. The value of each variable is given as the number of standard deviations above the mean the student falls (i.e.,

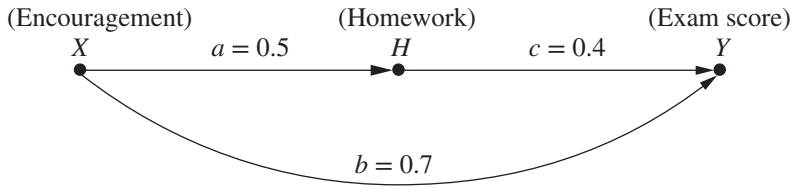

Figure 4.1 A model depicting the effect of Encouragement (X) on student's score

the model is *standardized* so that all variables have mean 0 and variance 1). For example, if $Y = 1$, then the student scored 1 standard deviation above the mean on his or her exam. This model represents a randomized pilot program, in which students are assigned to the remedial sessions by the luck of the draw.

Model 4.1

$$\begin{aligned} X &= U_X \\ H &= a \cdot X + U_H \\ Y &= b \cdot X + c \cdot H + U_Y \\ \sigma_{U_i U_j} &= 0 \quad \text{for all } i, j \in \{X, H, Y\} \end{aligned}$$

We assume that all U factors are independent and that we are given the values for the coefficients of Model 4.1 (these can be estimated from population data):

$$a = 0.5, \quad b = 0.7, \quad c = 0.4$$

Let us consider a student named Joe, for whom we measure $X = 0.5, H = 1,$ and $Y = 1.5$. Suppose we wish to answer the following query: What would Joe's score have been had he doubled his study time?

In a linear SEM, the value of each variable is determined by the coefficients and the U variables; the latter account for all variation among individuals. As a result, we can use the evidence $X = 0.5, H = 1,$ and $Y = 1.5$ to determine the values of the U variables associated with Joe. These values are invariant to hypothetical actions (or "miracles") such as those that might cause Joe to double his homework.

In this case, we are able to obtain the specific characteristics of Joe from the evidence:

$$\begin{aligned} U_X &= 0.5, \\ U_H &= 1 - 0.5 \cdot 0.5 = 0.75, \text{ and} \\ U_Y &= 1.5 - 0.7 \cdot 0.5 - 0.4 \cdot 1 = 0.75. \end{aligned}$$

Next, we simulate the action of doubling Joe's study time by replacing the structural equation for H with the constant $H = 2$. The modified model is depicted in Figure 4.2. Finally, we compute the value of Y in our modified model using the updated U values, giving

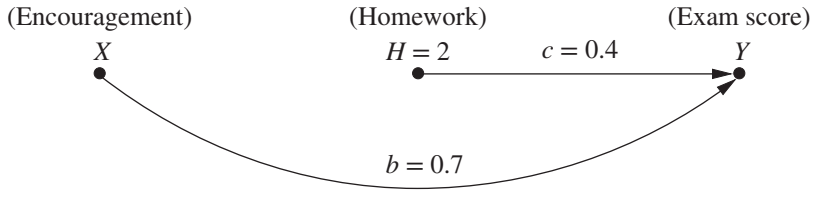

Figure 4.2 Answering a counterfactual question about a specific student's score, predicated on the assumption that homework would have increased to $H = 2$

$$\begin{aligned} Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) \\ &= 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 \\ &= 1.90 \end{aligned}$$

We thus conclude that Joe's score, had he doubled his homework, would have been 1.9 instead of 1.5. This, according to our convention, would mean an increase to 1.9 standard deviations above the mean, instead of the current 1.5.

In summary, we first applied the evidence $X = 0.5, H = 1$, and $Y = 1.5$ to update the values for the U variables. We then simulated an external intervention to force the condition $H = 2$ by replacing the structural equation $H = aX + U_H$ with the equation $H = 2$. Finally, we computed the value of Y given the structural equations and the updated U values. (In all of the above, we, of course, assumed that the U variables are unchanged by the hypothetical intervention on H.)

4.2.4 The Three Steps in Computing Counterfactuals

The case of Joe and the after-school program illustrates the way in which the fundamental definition of counterfactuals can be turned into a process for obtaining the value of a given counterfactual. There is a three-step process for computing any deterministic counterfactual:

(i) Abduction: Use evidence $E = e$ to determine the value of U.
(ii) Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
(iii) Prediction: Use the modified model, M_x, and the value of U to compute the value of Y, the consequence of the counterfactual.

In temporal metaphors, Step (i) explains the past (U) in light of the current evidence e; Step (ii) bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step (iii) predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.

This process will solve any deterministic counterfactual, that is, counterfactuals pertaining to a single unit of the population in which we know the value of every relevant variable. Structural equation models are able to answer counterfactual queries of this nature because each equation represents the mechanism by which a variable obtains its values. If we know these mechanisms, we should also be able to predict what values would be obtained had some of these mechanisms been altered, given the alterations. As a result, it is natural to view counterfactuals as *derived* properties of structural equations. (In some frameworks, counterfactuals are taken as primitives (Holland 1986; Rubin 1974).)

But counterfactuals can also be probabilistic, pertaining to a class of units within the population; for instance, in the after-school program example, we might want to know what would have happened if all students for whom $Y < 2$ had doubled their homework time. These probabilistic counterfactuals differ from *do*-operator interventions because, like their deterministic counterparts, they restrict the set of individuals intervened upon, which *do*-expressions cannot do.

We can now advance from deterministic to probabilistic models, so we can deal with questions about probabilities and expectations of counterfactuals. For example, suppose Joe is a student participating in the study of Figure 4.1, who scored $Y = y$ in the exam. What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement

training? Or, what would his *expected* score be in such hypothetical world? Unlike in the example of Model 4.1, we now do not have information on all three variables, $\{X, Y, H\}$, and we cannot therefore determine uniquely the value u that pertains to Joe. Instead, Joe may belong to a large class of units compatible with the evidence available, each having a different value of u.

Nondeterminism enters causal models by assigning probabilities $P(U = u)$ over the exogenous variables U. These represent our uncertainty as to the identity of the subject under consideration or, when the subject is known, what other characteristics that subject has that might have bearing on our problem.

The exogenous probability $P(U = u)$ induces a unique probability distribution on the endogenous variables $V, P(v)$, with the help of which we can define and compute not only the probability of any single counterfactual, $Y_x = y$, but also the joint distributions of all combinations of observed and counterfactual variables. For example, we can determine $P(Y_x = y, Z_w = z, X = x')$, where X, Y, Z, and W are arbitrary variables in a model. Such joint probabilities refer to the proportion of individuals u in the population for which all the events in the parentheses are true, namely, $Y_x(u) = y$ and $Z_w(u) = z$ and $X(u) = x'$, allowing, in particular, w or x' to conflict with x.

A typical query about these probabilities asks, "Given that we observe feature $E = e$ for a given individual, what would we expect the value of Y for that individual to be if X had been x?" This expectation is denoted $E[Y_{X=x}|E = e]$, where we allow $E = e$ to conflict with the antecedent $X = x$. $E = e$ after the conditioning bar represents all information (or evidence) we might have about the individual, potentially including the values of X, Y, or any other variable, as we have seen in Eq. (4.1). The subscript $X = x$ represents the antecedent specified by the counterfactual sentence.

The specifics of how these probabilities and expectations are dealt with will be examined in the following sections, but for now, it is important to know that using them, we can generalize our three-step process to any probabilistic nonlinear system.

Given an arbitrary counterfactual of the form, $E[Y_{X=x}|E =e]$, the three-step process reads:

(i) **Abduction**: Update $P(U)$ by the evidence to obtain $P(U|E = e)$.
(ii) **Action**: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
(iii) **Prediction**: Use the modified model, M_x, and the updated probabilities over the U variables, $P(U|E = e)$, to compute the expectation of Y, the consequence of the counterfactual.

We shall see in Section 4.4 that the above probabilistic procedure applies not only to retrospective counterfactual queries (queries of the form "What would have been the value of Y had X been x?") but also to certain kinds of intervention queries. In particular, it applies when we make every individual take an action that depends on the current value of his/her X. A typical example would be "additive intervention": for example, adding 5 mg/l of insulin to every patient's regiment, regardless of their previous dosage. Since the final level of insulin varies from patient to patient, this policy cannot be represented in *do*-notation.

For another example, suppose we wish to estimate, using Figure 4.1, the effect on test score provided by a school policy that sends students who are lazy on their homework ($H \leq H_0$) to attend the after-school program for $X = 1$. We can't simply intervene on X to set it equal to 1 in cases where H is low, because in our model, X is one of the causes of H.

Instead, we express the expected value of this quantity in counterfactual notation as $E[Y_{X=1}|H \leq H_0]$, which can, in principle, be computed using the above three-step method. Counterfactual reasoning and the above procedure are necessary for estimating the effect of actions and policies on subsets of the population characterized by features that, in themselves, are affected by the policy (e.g., $H \leq H_0$).

4.3 Nondeterministic Counterfactuals

4.3.1 Probabilities of Counterfactuals

To examine how nondeterminism is reflected in the calculation of counterfactuals, let us assign probabilities to the values of U in the model of Eqs. (4.3) and (4.4). Imagine that $U = \{1, 2, 3\}$ represents three types of individuals in a population, occurring with probabilities

$$P(U = 1) = \frac{1}{2}, P(U = 2) = \frac{1}{3}, \quad \text{and} \quad P(U = 3) = \frac{1}{6}$$

All individuals within a population type have the same values of the counterfactuals, as specified by the corresponding rows in Table 4.1. With these values, we can compute the probability that the counterfactuals will satisfy a specified condition. For instance, we can compute the proportion of units for which Y would be 3 had X been 2, or $Y_2(u) = 3$. This condition occurs only in the first row of the table and, since it is a property of $U = 1$, we conclude that it will occur with probability $\frac{1}{2}$, giving $P(Y_2 = 3) = \frac{1}{2}$. We can similarly compute the probability of any counterfactual statement, for example, $P(Y_1 = 4) = \frac{1}{6}, P(Y_1 = 3) = \frac{1}{3}, P(Y_2 > 3) = \frac{1}{2}$, and so on. What is remarkable, however, is that we can also compute joint probabilities of every combination of counterfactual and observable events. For example,

$$P(Y_2 > 3, Y_1 < 4) = \frac{1}{3}$$
$$P(Y_1 < 4, Y - X > 1) = \frac{1}{3}$$
$$P(Y_1 < Y_2) = 1$$

In the first of these expressions, we find a joint probability of two events occurring in two different worlds; the first $Y_2 > 3$ in an $X = 2$ world, and the second $Y_1 < 4$, in $X = 1$. The probability of their conjunction evaluates to $\frac{1}{3}$ because the two events co-occur only at $U = 2$, which was assigned a probability of $\frac{1}{3}$. Other cross-world events appear in the second and third expressions. Remarkably (and usefully), this clash between the worlds provides no barrier to calculation. In fact, cross-world probabilities are as simple to derive as intra-world ones: We simply identify the rows in which the specified combination is true and sum up the probabilities assigned to those rows. This immediately gives us the capability of computing conditional probabilities among counterfactuals and defining notions such as dependence and conditional independence among counterfactuals, as we did in Chapter 1 when we dealt with observable variables. For instance, it is easy to verify that, among individuals for which Y is greater than 2, the probability is $\frac{2}{3}$ that Y would increase if X were 3. (Because $P(Y_3 > Y|Y > 2) = \frac{1}{3}/\frac{1}{2} = \frac{2}{3}$.) Similarly, we can verify that the difference $Y_{x+1} - Y_x$ is independent of x, which means that the

causal effect of X on Y does not vary across population types, a property shared by all linear models.

Such joint probabilities over multiple-world counterfactuals can easily be expressed using the subscript notation, as in $P(Y_1 = y_1, Y_2 = y_2)$, and can be computed from any structural model as we did in Table 4.1. They cannot however be expressed using the $do(x)$ notation, because the latter delivers just one probability for each intervention $X = x$. To see the ramifications of this limitation, let us examine a slight modification of the model in Eqs. (4.3) and (4.4), in which a third variable Z acts as mediator between X and Y. The new model's equations are given by

$$X = U_1 \quad Z = aX + U_2, Y = bZ \tag{4.7}$$

and its structure is depicted in Figure 4.3. To cast this model in a context, let $X = 1$ stand for having a college education, $U_2 = 1$ for having professional experience, Z for the level of skill needed for a given job, and Y for salary.

Suppose our aim is to compute $E[Y_{X=1}|Z = 1]$, which stands for the expected salary of individuals with skill level $Z = 1$, had they received a college education. This quantity cannot be captured by a *do*-expression, because the condition $Z = 1$ and the antecedent $X = 1$ refer to two different worlds; the former represents current skills, whereas the latter represents a hypothetical education in an unrealized past. An attempt to capture this hypothetical salary using the expression $E[Y|do(X = 1), Z = 1]$ would not reveal the desired information. The *do*-expression stands for the expected salary of individuals who all finished college and have since acquired skill level $Z = 1$. The salaries of these individuals, as the graph shows, depend only on their skill, and are not affected by whether they obtained the skill through college or through work experience. Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in. In contrast, some of those who currently have $Z = 1$ might not have gone to college and would have attained higher skill (and salary) had they gotten college education. Their salaries are of great interest to us, but they are not included in the *do*-expression. Thus, in general, the *do*-expression will not capture our counterfactual question:

$$E[Y|do(X = 1),\ Z = 1] \neq E[Y_{X=1}|Z = 1] \tag{4.8}$$

We can further confirm this inequality by noting that, while $E[Y|do(X = 1), Z = 1]$ is equal to $E[Y|do(X = 0), Z = 1]$, $E[Y_{X=1}|Z = 1]$ is not equal to $E[Y_{X=0}|Z = 1]$; the formers treat $Z = 1$ as a postintervention condition that prevails for two different sets of units under the two antecedents, whereas the latters treat it as defining *one* set of units in the current world that would react differently under the two antecedents. The $do(x)$ notation cannot capture the latters because the events $X = 1$ and $Z = 1$ in the expression $E[Y_{X=1}|Z = 1]$ refer to two different worlds, pre- and postintervention, respectively. The expression $E[Y|do(X = 1), Z = 1]$

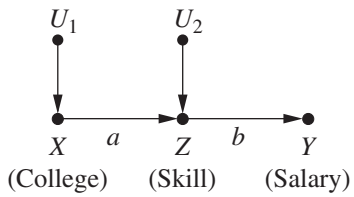

Figure 4.3 A model representing Eq. (4.7), illustrating the causal relations between college education (X), skills (Z), and salary (Y)

on the other hand, invokes only postintervention events, and that is why it is expressible in $do(x)$ notation.

A natural question to ask is whether counterfactual notation can capture the postintervention, single-world expression $E[Y|do(X = 1), Z = 1]$. The answer is affirmative; being more flexible, counterfactuals can capture both single-world and cross-world probabilities. The translation of $E[Y|do(X = 1), Z = 1]$ into counterfactual notation is simply $E[Y_{X=1}|Z_{X=1} = 1]$, which explicitly designates the event $Z = 1$ as postintervention. The variable $Z_{X=1}$ stands for the value that Z would attain had X been 1, and this is precisely what we mean when we put $Z = z$ in a *do*-expression by Bayes' rule:

$$P(Y = y|do(X = 1), Z = z) = \frac{P(Y = y, Z = z|do(X = 1))}{P(Z = z|do(X = 1))}$$

This shows explicitly how the dependence of Z on X should be treated. In the special case where Z is a preintervention variable, as age was in our discussion of conditional interventions (Section 3.5) we have $Z_{X=1} = Z$, and we need not distinguish between the two. The inequality in (4.8) then turns into an equality.

Let's look at how this logic is reflected in the numbers. Table 4.2 depicts the counterfactuals associated with the model of (4.7), with all subscripts denoting the state of X. It was constructed by the same method we used in constructing Table 4.1: replacing the equation $X = u$ with the appropriate constant (zero or one) and solving for Y and Z. Using this table, we can verify immediately that (see footnote 2)

$$E[Y_1|Z = 1] = (a + 1)b \tag{4.9}$$

$$E[Y_0|Z = 1] = b \tag{4.10}$$

$$E[Y|do(X = 1), Z = 1] = b \tag{4.11}$$

$$E[Y|do(X = 0), Z = 1] = b \tag{4.12}$$

These equations provide numerical confirmation of the inequality in (4.8). They also demonstrate a peculiar property of counterfactual conditioning that we have noted before: Despite the fact that Z separates X from Y in the graph of Figure 4.3, we find that X has an effect on Y for those units falling under $Z = 1$:

$$E[Y_1 - Y_0|Z = 1] = ab \neq 0$$

The reason for this behavior is best explained in the context of our salary example. While the salary of those who have acquired skill level $Z = 1$ depends only on their skill, not on X, the

Table 4.2 The values attained by $X(u), Y(u), Z(u), Y_0(u), Y_1(u), Z_0(u)$, and $Z_1(u)$ in the model of Eq. (4.7)

$X = u_1 \;\; Z = aX + u_2 \;\; Y = bZ$								
u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a + 1)b$	1	$a + 1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a + 1$	$(a + 1)b$	b	$(a + 1)b$	1	$a + 1$

[2]Strictly speaking, the quantity $E[Y|\, do(X{=}1), Z = 1]$ in Eq. (4.11) is undefined because the observation $Z = 1$ is not possible post-intervention of $do(X = 1)$. However, for the purposes of the example, we can imagine that $Z = 1$ was observed due to some error term $\varepsilon_z \rightarrow Z$ that accounts for the deviation. Eq. (4.11) then follows.

salary of those who are currently at $Z = 1$ would have been different *had they had a different past*. Retrospective reasoning of this sort, concerning dependence on the unrealized past, is not shown explicitly in the graph of Figure 4.3. To facilitate such reasoning, we need to devise means of representing counterfactual variables directly in the graph; we provide such representations in Section 4.3.2.

Thus far, the relative magnitudes of the probabilities of $P(u_1)$ and $P(u_2)$ have not entered into the calculations, because the condition $Z = 1$ occurs only for $u_1 = 0$ and $u_2 = 1$ (assuming that $a \neq 0$ and $a \neq 1$), and under these conditions, each of Y, Y_1, and Y_0 has a definite value. These probabilities play a role, however, if we assume $a = 1$ in the model, since $Z = 1$ can now occur under two conditions: $(u_1 = 0, u_2 = 1)$ and $(u_1 = 1, u_2 = 0)$. The first occurs with probability $P(u_1 = 0)P(u_2 = 1)$ and the second with probability $P(u_1 = 1)P(u_2 = 0)$. In such a case, we obtain

$$E[Y_{X=1}|Z = 1] = b\left(1 + \frac{P(u_1 = 0)P(u_2 = 1)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)}\right) \tag{4.13}$$

$$E[Y_{X=0}|Z = 1] = b\left(\frac{P(u_1 = 0)P(u_2 = 1)}{P(u_1 = 0)P(u_2 = 1) + P(u_1 = 1)P(u_2 = 0)}\right) \tag{4.14}$$

The fact that the first expression is larger than the second demonstrates again that the skill-specific causal effect of education on salary is nonzero, despite the fact that salaries are determined by skill only, not by education. This is to be expected, since a nonzero fraction of the workers at skill level $Z = 1$ did not receive college education, and, had they been given college education, their skill would have increased to $Z_1 = 2$, and their salaries to $2b$.

Study question 4.3.1

Consider the model in Figure 4.3 and assume that U_1 and U_2 are two independent Gaussian variables, each with zero mean and unit variance.

(a) Find the expected salary of workers at skill level $Z = z$ had they received x years of college education. [*Hint: Use Theorem 4.3.2, with $e : Z = z$, and the fact that for any two Gaussian variables, say X and Z, we have $E[X|Z = z] = E[X] + R_{XZ}(z - E[Z])$. Use the material in Sections 3.8.2 and 3.8.3 to express all regression coefficients in terms of structural parameters, and show that $E[Y_x|Z = z] = abx + bz/(1 + a^2)$.*]

(b) Based on the solution for (a), show that the skill-specific effect of education on salary is independent of the skill level.

4.3.2 The Graphical Representation of Counterfactuals

Since counterfactuals are byproducts of structural equation models, a natural question to ask is whether we can see them in the causal graphs associated with those models. The answer is affirmative, as can be seen from the fundamental law of counterfactuals, Eq. (4.5). This law tells us that if we modify model M to obtain the submodel M_x, then the outcome variable Y in the modified model is the counterfactual Y_x of the original model. Since modification calls for removing all arrows entering the variable X, as illustrated in Figure 4.4, we conclude that the node associated with the Y variable serves as a surrogate for Y_x, with the understanding that the substitution is valid only under the modification.

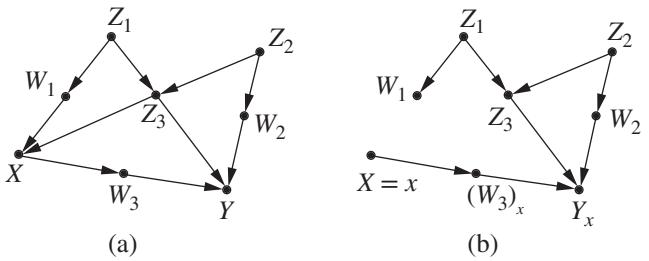

Figure 4.4 Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome Y predicated on $X = x$

This temporary visualization of counterfactuals is sufficient to answer some fundamental questions about the statistical properties of Y_x and how those properties depend on other variables in the model, specifically when those other variables are conditioned on.

When we ask about the statistical properties of Y_x, we need to examine what would cause Y_x to vary. According to its structural definition, Y_x represents the value of Y under a condition where X is held constant at $X = x$. Statistical variations of Y_x are therefore governed by all exogenous variables capable of influencing Y when X is held constant, that is, when the arrows entering X are removed, as in Figure 4.4(b). Under such conditions, the set of variables capable of transmitting variations to Y are the parents of Y (observed and unobserved), as well as parents of nodes on the pathways between X and Y. In Figure 4.4(b), for example, these parents are $\{Z_3, W_2, U_3, U_Y\}$, where U_Y and U_3, the error terms of Y and W_3, are not shown in the diagram. (These variables remain the same in both models.) Any set of variables that blocks a path to these parents also blocks that path to Y_x, and will result in, therefore, a conditional independence for Y_x. In particular, if we have a set Z of covariates that satisfies the backdoor criterion in M (see Definition 3.3.1), that set also blocks all paths between X and those parents, and consequently, it renders X and Y_x independent in every stratum $Z = z$.

These considerations are summarized formally in Theorem 4.3.1.

Theorem 4.3.1 (Counterfactual Interpretation of Backdoor) *If a set Z of variables satisfies the backdoor condition relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z*

$$P(Y_x|X, Z) = P(Y_x|Z) \tag{4.15}$$

Theorem 4.3.1 has far-reaching consequences when it comes to estimating the probabilities of counterfactuals from observational studies. In particular, it implies that $P(Y_x = y)$is identifiable by the adjustment formula of Eq. (3.5). To prove this, we condition on Z (as in Eq. (1.9)) and write

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y|Z = z)P(z) \\ &= \sum_z P(Y_x = y|Z = z, X = x)P(z) \\ &= \sum_z P(Y = y|Z = z, X = x)P(z) \end{aligned} \tag{4.16}$$

The second line was licensed by Theorem 4.3.1, whereas the third line was licensed by the consistency rule (4.6).

The fact that we obtained the familiar adjustment formula in Eq. (4.16) is not really surprising, because this same formula was derived in Section 3.2 (Eq. (3.4)), for $P(Y = y|do(x))$, and we know that $P(Y_x = y)$ is just another way of writing $P(Y = y|do(x))$. Interestingly, this derivation invokes only algebraic steps; it makes no reference to the model once we ensure that Z satisfies the backdoor criterion. Equation (4.15), which converts this graphical reality into algebraic notation, and allows us to derive (4.16), is sometimes called "conditional ignorability"; Theorem 4.3.1 gives this notion a scientific interpretation and permits us to test whether it holds in any given model.

Having a graphical representation for counterfactuals, we can resolve the dilemma we faced in Section 4.3.1 (Figure 4.3), and explain graphically why a stronger education (X) would have had an effect on the salary (Y) of people who are currently at skill level $Z = z$, despite the fact that, according to the model, salary is determined by skill only. Formally, to determine if the effect of education on salary (Y_x) is statistically independent of the level of education, we need to locate Y_x in the graph and see if it is d-separated from X given Z. Referring to Figure 4.3, we see that Y_x can be identified with U_2, the only parent of nodes on the causal path from X to Y (and therefore, the only variable that produces variations in Y_x while X is held constant). A quick inspection of Figure 4.3 tells us that Z acts as a collider between X and U_2, and, therefore, X and U_2 (and similarly X and Y_x) are not d-separated given Z. We conclude therefore

$$E[Y_x|X, Z] \neq E[Y_x|Z]$$

despite the fact that

$$E[Y|X, Z] = E[Y|Z]$$

In Study question 4.3.1, we evaluate these counterfactual expectations explicitly, assuming a linear Gaussian model. The graphical representation established in this section permits us to determine independencies among counterfactuals by graphical means, without assuming linearity or any specific parametric form. This is one of the tools that modern causal analysis has introduced to statistics, and, as we have seen in the analysis of the education–skill–salary story, it takes a task that is extremely hard to solve by unaided intuition and reduces it to simple operations on graphs. Additional methods of visualizing counterfactual dependencies, called "twin networks," are discussed in (Pearl 2000, pp. 213–215).

4.3.3 Counterfactuals in Experimental Settings

Having convinced ourselves that every counterfactual question can be answered from a fully specified structural model, we next move to the experimental setting, where a model is not available, and the experimenter must answer interventional questions on the basis of a finite sample of observed individuals. Let us refer back to the "encouragement design" model of Figure 4.1, in which we analyzed the behavior of an individual named Joe, and assume that the experimenter observes a set of 10 individuals, with Joe being participant 1. Each individual is characterized by a distinct vector $U_i = (U_X, U_H, U_Y)$, as shown in the first three columns of Table 4.3.

Table 4.3 Potential and observed outcomes predicted by the structural model of Figure 4.1 units were selected at random, with each U_i uniformly distributed over [0, 1]

	Participant characteristics			Observed behavior			Predicted potential outcomes				
Participant	U_X	U_H	U_Y	X	Y	H	Y_0	Y_1	H_0	H_1	$Y_{00}\cdots$
1	0.5	0.75	0.75	0.5	1.50	1.0	1.05	1.95	0.75	1.25	0.75
2	0.3	0.1	0.4	0.3	0.71	0.25	0.44	1.34	0.1	0.6	0.4
3	0.5	0.9	0.2	0.5	1.01	1.15	0.56	1.46	0.9	1.4	0.2
4	0.6	0.5	0.3	0.6	1.04	0.8	0.50	1.40	0.5	1.0	0.3
5	0.5	0.8	0.9	0.5	1.67	1.05	1.22	2.12	0.8	1.3	0.9
6	0.7	0.9	0.3	0.7	1.29	1.25	0.66	1.56	0.9	1.4	0.3
7	0.2	0.3	0.8	0.2	1.10	0.4	0.92	1.82	0.3	0.8	0.8
8	0.4	0.6	0.2	0.4	0.80	0.8	0.44	1.34	0.6	1.1	0.2
9	0.6	0.4	0.3	0.6	1.00	0.7	0.46	1.36	0.4	0.9	0.3
10	0.3	0.8	0.3	0.3	0.89	0.95	0.62	1.52	0.8	1.3	0.3

Using this information, we can create a full data set that complies with the model. For each triplet (U_X, U_H, U_Y), the model of Figure 4.1 enables us to complete a full row of the table, including Y_0 and Y_1, which stand for the potential outcomes under treatment ($X = 1$) and control ($X = 0$) conditions, respectively. We see that the structural model in Figure 4.1 encodes in effect a synthetic population of individuals together with their predicted behavior under both observational and experimental conditions. The columns labeled X,Y,H predict the results of observational studies, and those labeled Y_0, Y_1, H_0, H_1 predict the hypothetical outcome under two treatment regimes, $X = 0$, and $X = 1$. Many more, in fact infinite, potential outcomes may be predicted; for example, $Y_{X=0.5,Z=2.0}$ as computed for Joe from Figure 4.2, as well as all combinations of subscripted variables. From this synthetic population, one can estimate the probability of every counterfactual query on variables X,Y,H, assuming, of course, that we are in possession of all entries of the table. The estimation would require us to simply count the proportion of individuals that satisfy the specified query as demonstrated in Section 4.3.1.

Needless to say, the information conveyed by Table 4.3 is not available to us in either observational or experimental studies. This information was deduced from a parametric model such as the one in Figure 4.2, from which we could infer the defining characteristics $\{U_X,U_H,U_Y\}$ of each participant, given the observations $\{X,H,Y\}$. In general, in the absence of a parametric model, there is very little we learn about the potential outcomes Y_1 and Y_0 of individual participants, when all we have is their observed behavior $\{X,H,Y\}$. Theoretically, the only connection we have between the counterfactuals $\{Y_1,Y_0\}$ and the observables $\{X,H,Y\}$ is the consistency rule of Eq. (4.6), which informs us that Y_1 must be equal to Y in case X =1 and Y_0 must be equal to Y in case $X = 0$. But aside from this tenuous connection, most of the counterfactuals associated with the individual participants will remain unobserved.

Fortunately, there is much we can learn about those counterfactuals at the population level, such as estimating their probabilities or expectation. This we have witnessed already through the adjustment formula of (4.16), where we were able to compute $E(Y_1 - Y_0)$ using the graph

alone, instead of a complete model. Much more can be obtained from experimental studies, where even the graph becomes dispensable.

Assume that we have no information whatsoever about the underlying model. All we have are measurements on Y taken in an experimental study in which X is randomized over two levels, $X = 0$ and $X = 1$.

Table 4.4 describes the responses of the same 10 participants (Joe being participant 1) under such experimental conditions, with participants $1, 5, 6, 8$, and 10 assigned to $X = 0$, and the rest to $X = 1$. The first two columns give the true potential outcomes (taken from Table 4.3), while the last two columns describe the information available to the experimenter, where a square indicates that the response was not observed. Clearly, Y_0 is observed only for participants assigned to $X = 0$ and, similarly, Y_1 is observed only for those assigned to $X = 1$. Randomization assures us that, although half of the potential outcomes are not observed, the difference between the observed *means* in the treatment and control groups will converge to the difference of the population averages, $E(Y_1 - Y_0) = 0.9$. This is because randomization distributes the black squares at random along the two rightmost columns of Table 4.4, independent of the actual values of Y_0 and Y_1, so as the number of samples increases, the sample means converge to the population means.

This unique and important property of randomized experiments is not new to us, since randomization, like interventions, renders X independent of any variable that may affect Y (as in Figure 4.4(b)). Under such conditions, the adjustment formula (4.16) is applicable with $Z = \{\ \}$, yielding $E[Y_x] = E[Y|X = x]$, where $x = 1$ represents treated units and $x = 0$ untreated. Table 4.4 helps us understand what is actually computed when we take sample averages in experimental settings and how those averages are related to the underlying counterfactuals, Y_1 and Y_0.

Table 4.4 Potential and observed outcomes in a randomized clinical trial with X randomized over $X = 0$ and $X = 1$

	Predicted potential outcomes		Observed outcomes	
Participant	Y_0	Y_1	Y_0	Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■
	True average treatment effect: 0.90		Study average treatment effect: 0.68	

4.3.4 Counterfactuals in Linear Models

In nonparametric models, counterfactual quantities of the form $E[Y_{X=x}|Z = z]$ may not be identifiable, even if we have the luxury of running experiments. In fully linear models, however, things are much easier; any counterfactual quantity is identifiable whenever the model parameters are identified. This is because the parameters fully define the model's functions, and as we have seen earlier, once the functions are given, counterfactuals are computable using Eq. (4.5). Since every model parameter is identifiable from interventional studies using the interventional definition of direct effects, we conclude that in linear models, every counterfactual is experimentally identifiable. The question remains whether counterfactuals can be identified in observational studies, when some of the model parameters are not identified. It turns out that any counterfactual of the form $E[Y_{X=x}|Z = e]$, with e an arbitrary set of evidence, is identified whenever $E[Y|do(X = x)]$ is identified (Pearl 2000, p. 389). The relation between the two is summarized in Theorem 4.3.2, which provides a shortcut for computing counterfactuals.

Theorem 4.3.2 *Let τ be the slope of the total effect of X on Y,*

$$\tau = E[Y|do(x + 1)] - E[Y|do(x)]$$

then, for any evidence $Z = e$, we have

$$E[Y_{X=x}|Z = e] = E[Y|Z = e] + \tau(x - E[X|Z = e]) \tag{4.17}$$

This provides an intuitive interpretation of counterfactuals in linear models: $E[Y_{X=x}|Z = e]$ can be computed by first calculating the best estimate of Y conditioned on the evidence e, $E[Y|e]$, and then adding to it whatever change is expected in Y when X is shifted from its current best estimate, $E[X|Z = e]$, to its hypothetical value, x.

Methodologically, the importance of Theorem 4.3.2 lies in enabling researchers to answer hypothetical questions about individuals (or sets of individuals) from population data. The ramifications of this feature in legal and social contexts will be explored in the following sections. In the situation illustrated by Figure 4.2, we computed the counterfactual $Y_{H=2}$ under the evidence $e = \{X = 0.5, H = 1, Y = 1\}$. We now demonstrate how Theorem 4.3.2 can be applied to this model in computing the *effect of treatment on the treated*

$$ETT = E[Y_1 - Y_0|X = 1] \tag{4.18}$$

Substituting the evidence $e = \{X = 1\}$ in Eq. (4.17) we get

$$\begin{aligned} ETT &= E[Y_1|X = 1] - E[Y_0|X = 1] \\ &= E[Y|X = 1] - E[Y|X = 1] + \tau(1 - E[X|X = 1]) - \tau(0 - E[X|X = 1]) \\ &= \tau \\ &= b + ac = 0.9 \end{aligned}$$

In other words, the effect of treatment on the treated is equal to the effect of treatment on the entire population. This is a general result in linear systems that can be seen directly from Eq. (4.17); $E[Y_{x+1} - Y_x|e] = \tau$, independent on the evidence of e. Things are different when a

multiplicative (i.e., nonlinear) interaction term is added to the output equation. For example, if the arrow $X \rightarrow H$ were reversed in Figure 4.1, and the equation for Y read

$$Y = bX + cH + \delta XH + U_Y \tag{4.19}$$

τ would differ from ETT. We leave it to the reader as an exercise to show that the difference $\tau - ETT$ then equals $\frac{\delta a}{1+a^2}$ (see Study question 4.3.2(c)).

Study questions

Study question 4.3.2

(a) Describe how the parameters a, b, c in Figure 4.1 can be estimated from nonexperimental data.

(b) In the model of Figure 4.3, find the effect of education on those students whose salary is $Y = 1$. [Hint: Use Theorem 4.3.2 to compute $E[Y_1 - Y_0|Y = 1]$.]

(c) Estimate τ and the $ETT = E[Y_1 - Y_0|X = 1]$ for the model described in Eq. (4.19). [Hint: Use the basic definition of counterfactuals, Eq. (4.5) and the equality $E[Z|X = x'] = R_{ZX}x'$.]

4.4 Practical Uses of Counterfactuals

Now that we know how to compute counterfactuals, it will be instructive—and motivating—to see counterfactuals put to real use. In this section, we examine examples of problems that seem baffling at first, but that can be solved using the techniques we just laid out. Hopefully, the reader will leave this chapter with both a better understanding of how counterfactuals are used and a deeper appreciation of why we would want to use them.

4.4.1 Recruitment to a Program

Example 4.4.1 *A government is funding a job training program aimed at getting jobless people back into the workforce. A pilot randomized experiment shows that the program is effective; a higher percentage of people were hired among those who finished the program than among those who did not go through the program. As a result, the program is approved, and a recruitment effort is launched to encourage enrollment among the unemployed, by offering the job training program to any unemployed person who elects to enroll.*

Lo and behold, enrollment is successful, and the hiring rate among the program's graduates turns out even higher than in the randomized pilot study. The program developers are happy with the results and decide to request additional funding.

Oddly, critics claim that the program is a waste of taxpayers' money and should be terminated. Their reasoning is as follows: While the program was somewhat successful in the experimental study, where people were chosen at random, there is no proof that the program

accomplishes its mission among those who choose to enroll of their own volition. Those who self-enroll, the critics say, are more intelligent, more resourceful, and more socially connected than the eligible who did not enroll, and are more likely to have found a job regardless of the training. The critics claim that what we need to estimate is the *differential* benefit of the program on those enrolled: the extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not been trained.

Using our subscript notation for counterfactuals, and letting $X = 1$ represent training and $Y = 1$ represent hiring, the quantity that needs to be evaluated is the effect of training on the trained (ETT, better known as "effect of treatment on the treated," Eq. (4.18)):

$$ETT = E[Y_1 - Y_0|X = 1] \tag{4.20}$$

Here the difference $Y_1 - Y_0$ represents the causal effect of training (X) on hiring (Y) for a randomly chosen individual, and the condition $X = 1$ limits the choice to those actually choosing the training program on their own initiative.

As in our freeway example of Section 4.1, we are witnessing a clash between the antecedent ($X = 0$) of the counterfactual Y_0 (hiring had training not taken place) and the event it is conditioned on, $X = 1$. However, whereas the counterfactual analysis in the freeway example had no tangible consequences save for a personal regret statement—"I should have taken the freeway"—here the consequences have serious economic implications, such as terminating a training program, or possibly restructuring the recruitment strategy to attract people who would benefit more from the program offered.

The expression for ETT does not appear to be estimable from either observational or experimental data. The reason rests, again, in the clash between the subscript of Y_0 and the event $X = 1$ on which it is conditioned. Indeed, $E[Y_0|X = 1]$ stands for the expectation that a trained person ($X = 1$) would find a job had he/she not been trained. This counterfactual expectation seems to defy empirical measurement because we can never rerun history and deny training to those who received it. However, we see in the subsequent sections of this chapter that, despite this clash of worlds, the expectation $E[Y_0|X = 1]$ can be reduced to estimable expressions in many, though not all, situations. One such situation occurs when a set Z of covariates satisfies the backdoor criterion with regard to the treatment and outcome variables. In such a case, ETT probabilities are given by a modified adjustment formula:

$$\begin{aligned} &P(Y_x = y|X = x') \\ &\quad = \sum_z P(Y = y|X = x, Z = z)P(Z = z|X = x') \end{aligned} \tag{4.21}$$

This follows directly from Theorem 4.3.1, since conditioning on $Z = z$ gives

$$P(Y_x = y|x') = \sum_z P(Y_x = y|z, x')P(z|x')$$

but Theorem 4.3.1 permits us to replace x' with x, which by virtue of (4.6) permits us to remove the subscript x from Y_x, yielding (4.21).

Comparing (4.21) to the standard adjustment formula of Eq. (3.5),

$$P(Y = y|do(X = x)) = \sum P(Y = y|X = x, Z = z)P(Z = z)$$

we see that both formulas call for conditioning on $Z = z$ and averaging over z, except that (4.21) calls for a different weighted average, with $P(Z = z|X = x')$ replacing $P(Z = z)$.

Using Eq. (4.21), we readily get an estimable, noncounterfactual expression for ETT

$$\begin{aligned} ETT &= E[Y_1 - Y_0|X = 1] \\ &= E[Y_1|X = 1] - E[Y_0|X = 1] \\ &= E[Y|X = 1] - \sum_z E[Y|X = 0, Z = z]P(Z = z|X = 1) \end{aligned}$$

where the first term in the final expression is obtained using the consistency rule of Eq. (4.6). In other words, $E[Y_1|X = 1] = E[Y|X = 1]$ because, conditional on $X = 1$, the value that Y would get had X been 1 is simply the observed value of Y.

Another situation permitting the identification of ETT occurs for binary X whenever both experimental and nonexperimental data are available, in the form of $P(Y = y|do(X = x))$ and $P(X = x, Y = y)$, respectively. Still another occurs when an intermediate variable is available between X and Y satisfying the front-door criterion (Figure 3.10(b)). What is common to these situations is that an inspection of the causal graph can tell us whether ETT is estimable and, if so, how.

Study questions

Study question 4.4.1

(a) Prove that, if X is binary, the effect of treatment on the treated can be estimated from both observational and experimental data. Hint: Decompose $E[Y_x]$ into

$$E[Y_x] = E[Y_x|x']P(x') + E[Y_x|x]P(x)$$

(b) Apply the result of Question (a) to Simpson's story with the nonexperimental data of Table 1.1, and estimate the effect of treatment on those who used the drug by choice. [*Hint: Estimate $E[Y_x]$ assuming that gender is the only confounder.*]

(c) Repeat Question (b) using Theorem 4.3.1 and the fact that Z in Figure 3.3 satisfies the backdoor criterion. Show that the answers to (b) and (c) coincide.

4.4.2 Additive Interventions

Example 4.4.2 *In many experiments, the external manipulation consists of adding (or subtracting) some amount from a variable X without disabling preexisting causes of X, as required by the do(x) operator. For example, we might give 5 mg/l of insulin to a group of patients with varying levels of insulin already in their systems. Here, the preexisting causes of the manipulated variable continue to exert their influences, and a new quantity is added, allowing for differences among units to continue. Can the effect of such interventions be predicted from observational studies, or from experimental studies in which X was set uniformly to some predetermined value x?*

If we write our question using counterfactual variables, the answer becomes obvious. Suppose we were to add a quantity q to a treatment variable X that is currently at level $X = x'$.

The resulting outcome would be $Y_{x'+q}$, and the average value of this outcome over all units currently at level $X = x'$ would be $E[Y_x|x']$, with $x = x' + q$. Here, we meet again the ETT expression $E[Y_x|x']$, to which we can apply the results described in the previous example. In particular, we can conclude immediately that, whenever a set Z in our model satisfies the backdoor criterion, the effect of an additive intervention is estimable using the ETT adjustment formula of Eq. (4.21). Substituting $x = x' + q$ in (4.21) and taking expectations gives the effect of this intervention, which we call $add(q)$:

$$\begin{aligned} &E[Y|add(q)] - E[Y] \\ &\quad = \sum_{x'} E[Y_{x'+q}|X = x']P(X = x') - E[Y] \\ &\quad = \sum_{x'} \sum_{z} E[Y|X = x' + q, Z = z]P(Z = z|X = x')P(X = x') - E[Y] \end{aligned} \tag{4.22}$$

In our example, Z may include variables such as age, weight, or genetic disposition; we require only that each of those variables be measured and that they satisfy the backdoor condition.

Similarly, estimability is assured for all other cases in which ETT is identifiable.

This example demonstrates the use of counterfactuals to estimate the effect of practical interventions, which cannot always be described as *do*-expressions, but may nevertheless be estimated under certain circumstances. A question naturally arises: Why do we need to resort to counterfactuals to predict the effect of a rather common intervention, one that could be estimated by a straightforward clinical trial at the population level? We simply split a randomly chosen group of subjects into two parts, subject half of them to an $add(q)$ type of intervention and compare the expected value of Y in this group to that obtained in the $add(0)$ group. What is it about additive interventions that force us to seek the advice of a convoluted oracle, in the form of counterfactuals and ETT, when the answer can be obtained by a simple randomized trial?

The answer is that we need to resort to counterfactuals only because our target quantity, $E[Y|add(q)]$, could not be reduced to *do*-expressions, and it is through *do*-expressions that scientists report their experimental findings. This does not mean that the desired quantity $E[Y|add(q)]$ cannot be obtained from a specially designed experiment; it means only that save for conducting such a special experiment, the desired quantity cannot be inferred from scientific knowledge or from a standard experiment in which X is set to $X = x$ uniformly over the population. The reason we seek to base policies on such ideal standard experiments is that they capture scientific knowledge. Scientists are interested in quantifying the effect of increasing insulin concentration in the blood from a given level $X = x$ to a another level $X = x + q$, and this increase is captured by the *do*-expression: $E[Y|do(X = x + q)] - E[Y|do(X = x)]$. We label it "scientific" because it is biologically meaningful, namely its implications are invariant across populations (indeed laboratory blood tests report patients' concentration levels, $X = x$, which are tracked over time). In contrast, the policy question in the case of additive interventions does not have this invariance feature; it asks for the average effect of adding an increment q to everyone, regardless of the current x level of each individual in this particular population. It is not immediately transportable, because it is highly sensitive to the probability $P(X = x)$ in the population under study. This creates a mismatch between what science tells us and what policy makers ask us to estimate. It is no wonder, therefore, that we need to resort to a unit-level analysis (i.e., counterfactuals) in order to translate from one language into another.

The reader may also wonder why $E[Y|add(q)]$ is not equal to the average causal effect

$$\sum_x [E[Y|do(X = x + q)] - E[Y|do(X = x)]]\ P(X = x)$$

After all, if we know that adding q to an individual at level $X = x$ would increase its expected Y by $E[Y|do(X = x + q)] - E[Y|do(X = x)]$, then averaging this increase over X should give us the answer to the policy question $E[Y|add(q)]$. Unfortunately, this average does *not* capture the policy question. This average represents an experiment in which subjects are chosen at random from the population, a fraction $P(X = x)$ are given an additional dose q, and the rest are left alone. But things are different in the policy question at hand, since $P(X = x)$ represents the proportion of subjects who entered level $X = x$ by free choice, and we cannot rule out the possibility that subjects who attain $X = x$ by free choice would react to $add(q)$ differently from subjects who "receive" $X = x$ by experimental decree. For example, it is quite possible that subjects who are highly sensitive to $add(q)$ would attempt to lower their X level, given the choice.

We translate into counterfactual analysis and write the inequality:

$$E[Y|add(q)] = \sum_x E[Y_{x+q}|x]P(X = x) \neq \sum_x E[Y_{x+q}]P(X = x)$$

Equality holds only when Y_x is independent of X, a condition that amounts to nonconfounding (see Theorem 4.3.1). Absent this condition, the estimation of $E[Y|add(q)]$ can be accomplished either by q-specific intervention or through stronger assumptions that enable the translation of ETT to *do*-expressions, as in Eq. (4.21).

Study question 4.4.2

Joe has never smoked before but, as a result of peer pressure and other personal factors, he decided to start smoking. He buys a pack of cigarettes, comes home, and asks himself: "I am about to start smoking, should I?"

(a) Formulate Joe's question mathematically, in terms of ETT, assuming that the outcome of interest is lung cancer.
(b) What type of data would enable Joe to estimate his chances of getting cancer given that he goes ahead with the decision to smoke, versus refraining from smoking.
(c) Use the data in Table 3.1 to estimate the chances associated with the decision in (b).

4.4.3 Personal Decision Making

Example 4.4.3 *Ms Jones, a cancer patient, is facing a tough decision between two possible treatments: (i) lumpectomy alone or (ii) lumpectomy plus irradiation. In consultation with her oncologist, she decides on (ii). Ten years later, Ms Jones is alive, and the tumor has not recurred. She speculates: Do I owe my life to irradiation?*

Mrs Smith, on the other hand, had a lumpectomy alone, and her tumor recurred after a year. And she is regretting: I should have gone through irradiation.

Can these speculations ever be substantiated from statistical data? Moreover, what good would it do to confirm Ms Jones's triumph or Mrs Smith's regret?

The overall effectiveness of irradiation can, of course, be determined by randomized experiments. Indeed, on October 17, 2002, the *New England Journal of Medicine* published a paper by Fisher et al. describing a 20-year follow-up of a randomized trial comparing lumpectomy alone and lumpectomy plus irradiation. The addition of irradiation to lumpectomy was shown to cause substantially fewer recurrences of breast cancer (14% vs 39%).

These, however, were population results. Can we infer from them the specific cases of Ms Jones and Mrs Smith? And what would we gain if we do, aside from supporting Ms Jones's satisfaction with her decision or intensifying Mrs Smith's sense of failure?

To answer the first question, we must first cast the concerns of Ms Jones and Mrs Smith in mathematical form, using counterfactuals. If we designate remission by $Y = 1$ and the decision to undergo irradiation by $X = 1$, then the probability that determines whether Ms Jones is justified in attributing her remission to the irradiation ($X = 1$) is

$$PN = P(Y_0 = 0|X = 1, Y = 1) \tag{4.23}$$

It reads: the probability that remission would *not* have occurred ($Y = 0$) had Ms Jones not gone through irradiation, given that she did in fact go through irradiation ($X = 1$), and remission did occur ($Y = 1$). The label PN stands for "probability of necessity" that measures the degree to which Ms Jones's decision was *necessary* for her positive outcome.

Similarly, the probability that Ms Smith's regret is justified is given by

$$PS = P(Y_1 = 1|X = 0, Y = 0) \tag{4.24}$$

It reads: the probability that remission would have occurred had Mrs Smith gone through irradiation ($Y_1 = 1$), given that she did not in fact go through irradiation ($X = 0$), and remission did not occur ($Y = 0$). PS stands for the "probability of sufficiency," measuring the degree to which the action $X = 1$, which was not taken.

We see that these expressions have almost the same form (save for interchanging ones with zeros) and, moreover, both are similar to Eq. (4.1), save for the fact that Y in the freeway example was a continuous variable, so its expected value was the quantity of interest.

These two probabilities (sometimes referred to as "probabilities of causation") play a major role in all questions of "attribution," ranging from legal liability to personal decision making. They are not, in general, estimable from either observational or experimental data, but as we shall see below, they *are* estimable under certain conditions, when both observational and experimental data are available.

But before commencing a quantitative analysis, let us address our second question: What is gained by assessing these retrospective counterfactual parameters? One answer is that notions such as regret and success, being right or being wrong, have more than just emotional value; they play important roles in cognitive development and adaptive learning. Confirmation of Ms Jones's triumph reinforces her confidence in her decision-making strategy, which may include her sources of medical information, her attitude toward risks, and her sense of priority, as well as the strategies she has been using to put all these considerations together. The same applies to regret; it drives us to identify sources of weakness in our strategies and to think of some kind of change that would improve them. It is through counterfactual reinforcement that we learn to improve our own decision-making processes and achieve higher performance. As Kathryn Schultz says in her delightful book *Being Wrong*, "However disorienting, difficult, or humbling our mistakes might be, it is ultimately wrongness, not rightness, that can teach us who we are."

Estimating the probabilities of being right or wrong also has tangible and profound impact on critical decision making. Imagine a third lady, Ms Daily, facing the same decision as Ms Jones did, and telling herself: If my tumor is the type that would not recur under lumpectomy alone, why should I go through the hardships of irradiation? Similarly, if my tumor is the type that would recur regardless of whether I go through irradiation or not, I would rather not go through it. The only reason for me to go through this is if the tumor is the type that would remiss under treatment and recur under no treatment.

Formally, Ms Daily's dilemma is to quantify the probability that irradiation is both *necessary and sufficient* for eliminating her tumor, or

$$PNS = P(Y_1 = 1, Y_0 = 0) \tag{4.25}$$

where Y_1 and Y_0 stand for remission under treatment (Y_1) and nontreatment (Y_0), respectively. Knowing this probability would help Ms Daily's assessment of how likely she is to belong to the group of individuals for whom $Y_1 = 1$ and $Y_0 = 0$.

This probability cannot, of course, be assessed from experimental studies, because we can never tell from experimental data whether an outcome would have been different had the person been assigned to a different treatment. However, casting Ms Daily's question in mathematical form enables us to investigate algebraically what assumptions are needed for estimating PNS and from what type of data. In the next section (Section 4.5.1, Eq. (4.42)), we see that indeed, PNS can be estimated if we assume *monotonicity*, namely, that irradiation cannot cause the recurrence of a tumor that was about to remit. Moreover, under monotonicity, experimental data are sufficient to conclude

$$PNS = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{4.26}$$

For example, if we rely on the experimental data of Fisher et al. (2002), this formula permits us to conclude that Ms Daily's PNS is

$$PNS = 0.86 - 0.61 = 0.25$$

This gives her a 25% chance that her tumor is the type that responds to treatment—specifically, that it will remit under lumpectomy plus irradiation but will recur under lumpectomy alone. Such quantification of individual risks is extremely important in personal decision making, and estimates of such risks from population data can only be inferred through counterfactual analysis and appropriate assumptions.

4.4.4 Discrimination in Hiring

Example 4.4.4 *Mary files a law suit against the New York-based XYZ International, alleging discriminatory hiring practices. According to her, she has applied for a job with XYZ International, and she has all the credentials for the job, yet she was not hired, allegedly because she mentioned, during the course of her interview, that she is gay. Moreover, she claims, the hiring record of XYZ International shows consistent preferences for straight employees. Does she have a case? Can hiring records prove whether XYZ International was discriminating when declining her job application?*

At the time of writing, U.S. law doesn't specifically prohibit employment discrimination on the basis of sexual orientation, but New York law does. And New York defines discrimination in much the same way as federal law. U.S. courts have issued clear directives as to what constitutes employment discrimination. According to law makers, "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin, etc.) and everything else had been the same." (In Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).)

The first thing to note in this directive is that it is not a population-based criterion, but one that appeals to the individual case of the plaintiff. The second thing to note is that it is formulated in counterfactual terminology, using idioms such as "would have taken," "had the employee been," and "had been the same." What do they mean? Can one ever prove how an employer would have acted had Mary been straight? Certainly, this is not a variable that we can intervene upon in an experimental setting. Can data from an observational study prove an employer discriminating?

It turns out that Mary's case, though superficially different from Example 4.4.3, has a lot in common with the problem Ms Smith faced over her unsuccessful cancer treatment. The probability that Mary's nonhiring is due to her sexual orientation can, similarly to Ms Smith's cancer treatment, be expressed using the probability of sufficiency:

$$PS = P(Y_1 = 1 | X = 0, Y = 0)$$

In this case, Y stands for Mary's hiring, and X stands for the interviewer's perception of Mary's sexual orientation. The expression reads: "the probability that Mary would have been hired had the interviewer perceived her as straight, given that the interviewer perceived her as gay, and she was not hired." (Note that the variable in question is the interviewer's *perception* of Mary's sexual orientation, not the orientation itself, because an intervention on perception is quite simple in this case—we need only to imagine that Mary never mentioned that she is gay; hypothesizing a change in Mary's actual orientation, although formally acceptable, brings with it an aura of awkwardness.)

We show in 4.5.2 that, although discrimination cannot be proved in individual cases, the probability that such discrimination took place can be determined, and this probability may sometimes reach a level approaching certitude. The next example examines how the problem of discrimination—in this case on gender, not sexual orientation may appear to a policy maker, rather than a juror.

4.4.5 Mediation and Path-disabling Interventions

Example 4.4.5 *A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training. The former concerns the "direct effect" of gender on hiring, whereas the latter concerns the "indirect effect," or the effect* mediated *via job qualification.*

In this example, fighting employers' prejudices and launching educational reforms are two contending policy options that involve costly investments and different implementation strategies. Knowing in advance which of the two, if successful, would have a greater impact on reducing hiring disparity is essential for planning, and depends critically on mediation analysis for resolution. For example, knowing that current hiring disparities are due primarily to employers' prejudices would render educational reforms superfluous, a fact that may save substantial resources. Note, however, that the policy decisions in this example concern the enabling and disabling of processes rather than lowering or raising values of specific variables. The educational reform program calls for disabling current educational practices and replacing them with a new program in which women obtain the same educational opportunities as men. The hiring-based proposal calls for disabling the current hiring process and replacing it with one in which gender plays no role in hiring decisions.

Because we are dealing with disabling processes rather than changing levels of variables, there is no way we can express the effect of such interventions using a *do*-operator, as we did in the mediation analysis of Section 3.7. We can express it, however, in a counterfactual language, using the desired end result as an antecedent. For example, if we wish to assess the hiring disparity after successfully implementing gender-blind hiring procedures, we impose the condition that all female applicants be treated like males as an antecedent and proceed to estimate the hiring rate under such a counterfactual condition.

The analysis proceeds as follows: the hiring status (Y) of a female applicant with qualification $Q = q$, given that the employer treats her as though she is a male is captured by the counterfactual $Y_{X=1,Q=q}$, where $X = 1$ refers to being a male. But since the value q would vary among applicants, we need to average this quantity according to the distribution of female qualification, giving $\sum_q E[Y_{X=1,Q=q}]P(Q = q|X = 0)$. Male applicants would have a similar chance at hiring except that the average is governed by the distribution of male qualification, giving

$$\sum_q E[Y_{X=1,Q=q}]P(Q = q|X = 1)$$

If we subtract the two quantities, we get

$$\sum_q E[Y_{X=1,Q=q}][P(Q = q|X = 0) - P(Q = q|X = 1)]$$

which is the indirect effect of gender on hiring, mediated by qualification. We call this effect the natural indirect effect (NIE), because we allow the qualification Q to vary naturally from applicant to applicant, as opposed to the controlled direct effect in Chapter 3, where we held the mediator at a constant level for the entire population. Here we merely disable the capacity of Y to respond to X but leave its response to Q unaltered.

The next question to ask is whether such a counterfactual expression can be identified from data. It can be shown (Pearl 2001) that, in the absence of confounding the NIE can be estimated by conditional probabilities, giving

$$NIE = \sum_q E[Y|X = 1, Q = q][P(Q = q|X = 0) - P(Q = q|X = 1)]$$

This expression is known as the *mediation formula*. It measures the extent to which the effect of X on Y is *explained* by its effect on the mediator Q. Counterfactual analysis permits us to define and assess NIE by "freezing" the direct effect of X on Y, and allowing the mediator (Q) of each unit to react to X in a natural way, as if no freezing took place.

The mathematical tools necessary for estimating the various nuances of mediation are summarized in Section 4.5.

4.5 Mathematical Tool Kits for Attribution and Mediation

As we examined the practical applications of counterfactual analysis in Section 4.4, we noted several recurring patterns that shared mathematical expressions as well as methods of solution. The first was the effect of treatment on the treated, *ETT*, whose syntactic signature was the counterfactual expression $E[Y_x|X = x']$, with x and x' two distinct values of X. We showed that problems as varied as recruitment to a program (Section 4.4.1) and additive interventions (Example 4.4.2) rely on the estimation of this expression, and we have listed conditions under which estimation is feasible, as well as the resulting estimand (Eqs. (4.21) and (4.8)).

Another recurring pattern appeared in problems of attribution, such as personal decision problems (Example 4.4.3) and possible cases of discrimination (Example 4.4.4). Here, the pattern was the expression for the probability of necessity:

$$PN = P(Y_0 = 0|X = 1, Y = 1)$$

The probability of necessity also pops up in problems of legal liability, where it reads: "The probability that the damage would not have occurred had the action not been taken ($Y_0 = 0$), given that, in fact, the damage did occur ($Y = 1$) and the action was taken ($X = 1$)." Section 4.5.1 summarizes mathematical results that will enable readers to estimate (or bound) PN using a combination of observational and experimental data.

Finally, in questions of mediation (Example 4.4.5) the key counterfactual expression was

$$E[Y_{x,M_{x'}}]$$

which reads, "The expected outcome (Y) had the treatment been $X = x$ and, simultaneously, had the mediator M attained the value ($M_{x'}$) it would have attained had X been x'". Section 4.5.2 will list the conditions under which this "nested" counterfactual expression can be estimated, as well as the resulting estimands and their interpretations.

4.5.1 A Tool Kit for Attribution and Probabilities of Causation

Assuming binary events, with $X = x$ and $Y = y$ representing treatment and outcome, respectively, and $X = x'$, $Y = y'$ their negations, our target quantity is defined by the English sentence:

> "Find the probability that if X had been x', Y would be y', given that, in reality, X is x and Y is y."

Mathematically, this reads

$$PN(x, y) = P(Y_{x'} = y' | X = x, Y = y) \tag{4.27}$$

This counterfactual quantity, named "probability of necessity" (PN), captures the legal criterion of "but for," according to which judgment in favor of a plaintiff should be made if and only if it is "more probable than not" that the damage would not have occurred *but for* the defendant's action (Robertson 1997).

Having written a formal expression for PN, Eq. (4.27), we can move on to the identification phase and ask what assumptions permit us to identify PN from empirical studies, be they observational, experimental, or a combination thereof.

Mathematical analysis of this problem (described in (Pearl 2000, Chapter 9)) yields the following results:

Theorem 4.5.1 *If Y is monotonic relative to X, that is,* $Y_1(u) \geq Y_0(u)$ *for all u, then PN is identifiable whenever the causal effect* $P(y|do(x))$ *is identifiable, and*

$$PN = \frac{P(y) - P(y|do(x'))}{P(x, y)} \tag{4.28}$$

or, substituting $P(y) = P(y|x)P(x) + P(y|x')(1 - P(x))$*, we obtain*

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x, y)} \tag{4.29}$$

The first term on the r.h.s. of (4.29) is called the *excess risk ratio* (ERR) and is often used in court cases in the absence of experimental data (Greenland 1999). It is also known as the Attributable Risk Fraction among the exposed (Jewell 2004, Chapter 4.7). The second term (the *confounding factor* (CF)) represents a *correction* needed to account for confounding bias, that is, $P(y|do(x')) \neq P(y|x')$. Put in words, confounding occurs when the proportion of population for whom $Y = y$, when X is set to x' for everyone is not the same as the proportion of the population for whom $Y = y$ among those acquiring $X = x'$ by choice. For instance, suppose there is a case brought against a car manufacturer, claiming that its car's faulty design led to a man's death in a car crash. The ERR tells us how much more likely people are to die in crashes when driving one of the manufacturer's cars. If it turns out that people who buy the manufacturer's cars are more likely to drive fast (leading to deadlier crashes) than the general population, the second term will correct for that bias.

Equation (4.29) thus provides an estimable measure of necessary causation, which can be used for monotonic $Y_x(u)$ whenever the causal effect $P(y|do(x))$ can be estimated, be it from randomized trials or from graph-assisted observational studies (e.g., through the backdoor criterion). More significantly, it has also been shown (Tian and Pearl 2000) that the expression in (4.28) provides a lower bound for PN in the general nonmonotonic case. In particular, the upper and lower bounds on PN are given by

$$\max\left\{0, \frac{P(y) - P(y|do(x'))}{P(x, y)}\right\} \leq PN \leq \min\left\{1, \frac{P(y'|do(x')) - P(x', y')}{P(x, y)}\right\} \tag{4.30}$$

In drug-related litigation, it is not uncommon to obtain data from both experimental and observational studies. The former is usually available from the manufacturer or the agency

that approved the drug for distribution (e.g., FDA), whereas the latter is often available from surveys of the population.

A few algebraic steps allow us to express the lower bound (*LB*) and upper bound (*UB*) as

$$LB = ERR + CF$$
$$UB = ERR + q + CF \tag{4.31}$$

where *ERR*, *CF*, and q are defined as follows:

$$CF \triangleq [P(y|x') - P(y_{x'})]/P(x, y) \tag{4.32}$$
$$ERR \triangleq 1 - 1/RR = 1 - P(y|x')/P(y|x) \tag{4.33}$$
$$q \triangleq P(y'|x)/P(y|x) \tag{4.34}$$

Here, CF represents the normalized degree of confounding among the unexposed ($X = x'$), ERR is the "excess risk ratio" and q is the ratio of negative to positive outcomes among the exposed.

Figure 4.5(a) and (b) depicts these bounds as a function of ERR, and reveals three useful features. First, regardless of confounding, the interval *UB–LB* remains constant and depends on only one observable parameter, $P(y'|x)/P(y|x)$. Second, the CF may raise the lower bound to meet the criterion of "more probable than not," $PN > \frac{1}{2}$, when the ERR alone would not suffice. Lastly, the amount of "rise" to both bounds is given by *CF*, which is the only estimate needed from the experimental data; the causal effect $P(y_x) - P(y_{x'})$ is not needed.

Theorem 4.5.1 further assures us that, if monotonicity can be assumed, the upper and lower bounds coincide, and the gap collapses entirely, as shown in Figure 4.5(b). This collapse does not reflect $q = 0$, but a shift from the bounds of (4.30) to the identified conditions of (4.28).

If it is the case that the experimental and survey data have been drawn at random from the same population, then the experimental data can be used to estimate the counterfactuals

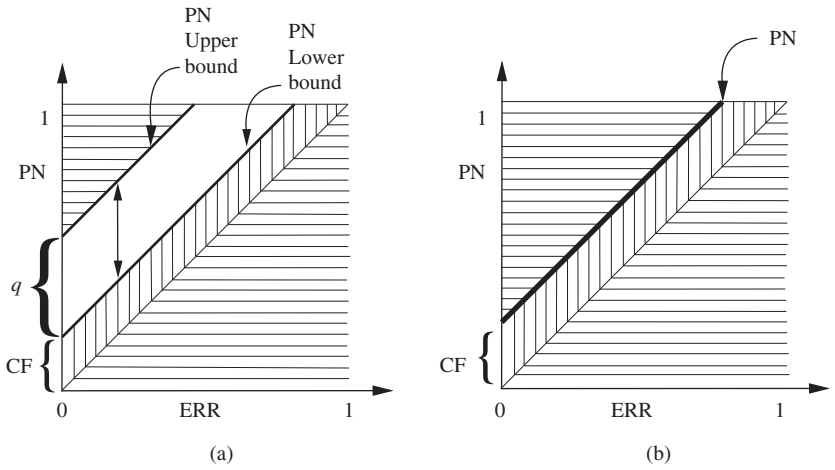

Figure 4.5 (a) Showing how probabilities of necessity (PN) are bounded, as a function of the excess risk ratio (ERR) and the confounding factor (CF) (Eq. (4.31)); (b) showing how PN is identified when monotonicity is assumed (Theorem 4.5.1)

of interest, for example, $P(Y_x = y)$, for the observational as well as experimental sampled populations.

Example 4.5.1 (Attribution in Legal Setting) *A lawsuit is filed against the manufacturer of drug x, charging that the drug is likely to have caused the death of Mr A, who took it to relieve back pains. The manufacturer claims that experimental data on patients with back pains show conclusively that drug x has only minor effects on death rates. However, the plaintiff argues that the experimental study is of little relevance to this case because it represents average effects on patients in the study, not on patients like Mr A who did not participate in the study. In particular, argues the plaintiff, Mr A is unique in that he used the drug of his own volition, unlike subjects in the experimental study, who took the drug to comply with experimental protocols. To support this argument, the plaintiff furnishes nonexperimental data on patients who, like Mr A, chose drug x to relieve back pains but were not part of any experiment, and who experienced lower death rates than those who didn't take the drug. The court must now decide, based on both the experimental and nonexperimental studies, whether it is "more probable than not" that drug x was in fact the cause of Mr A's death.*

To illustrate the usefulness of the bounds in Eq. (4.30), consider (hypothetical) data associated with the two studies shown in Table 4.5. (In the analyses below, we ignore sampling variability.)

The experimental data provide the estimates

$$P(y|do(x)) = 16/1000 = 0.016 \tag{4.35}$$

$$P(y|do(x')) = 14/1000 = 0.014 \tag{4.36}$$

whereas the nonexperimental data provide the estimates

$$P(y) = 30/2000 = 0.015 \tag{4.37}$$

$$P(x, y) = 2/2000 = 0.001 \tag{4.38}$$

$$P(y|x) = 2/1000 = 0.002 \tag{4.39}$$

$$P(y|x') = 28/1000 = 0.028 \tag{4.40}$$

Table 4.5 Experimental and nonexperimental data used to illustrate the estimation of PN, the probability that drug x was responsible for a person's death (y)

	Experimental		Nonexperimental	
	$do(x)$	$do(x')$	x	x'
Deaths (y)	16	14	2	28
Survivals (y')	984	986	998	972

Assuming that drug x can only cause (but never prevent) death, monotonicity holds, and Theorem 4.5.1 (Eq. (4.29)) yields

$$PN = \frac{P(y|x) - P(y|x')}{P(y|x)} + \frac{P(y|x') - P(y|do(x'))}{P(x,y)}$$
$$= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = -13 + 14 = 1 \tag{4.41}$$

We see that while the observational ERR is negative (-13), giving the impression that the drug is actually preventing deaths, the bias-correction term ($+14$) rectifies this impression and sets the probability of necessity (PN) to unity. Moreover, since the lower bound of Eq. (4.30) becomes 1, we conclude that $PN = 1.00$ even without assuming monotonicity. Thus, the plaintiff was correct; barring sampling errors, the data provide us with 100% assurance that drug x was in fact responsible for the death of Mr A.

To complete this tool kit for attribution, we note that the other two probabilities that came up in the discussion on personal decision-making (Example 4.4.3), PS and PNS, can be bounded by similar expressions; see (Pearl 2000, Chapter 9) and (Tian and Pearl 2000).

In particular, when $Y_x(u)$ is monotonic, we have

$$PNS = P(Y_x = 1, Y_{x'} = 0)$$
$$= P(Y_x = 1) - P(Y_{x'} = 1) \tag{4.42}$$

as asserted in Example 4.4.3, Eq. (4.26).

Study questions

Study question 4.5.1

Consider the dilemma faced by Ms Jones, as described in Example 4.4.3. Assume that, in addition to the experimental results of Fisher et al. (2002), she also gains access to an observational study, according to which the probability of recurrent tumor in all patients (regardless of irradiation) is 30%, whereas among the recurrent cases, 70% did not choose therapy. Use the bounds provided in Eq. (4.30) to update her estimate that her decision was necessary for remission.

4.5.2 A Tool Kit for Mediation

The canonical model for a typical mediation problem takes the form:

$$t = f_T(u_T) \quad m = f_M(t, u_M) \quad y = f_Y(t, m, u_Y) \tag{4.43}$$

where T (treatment), M (mediator), and Y (outcome) are discrete or continuous random variables, f_T,f_M, and f_Y are arbitrary functions, and U_T, U_M, U_Y represent, respectively, omitted factors that influence T, M, and Y. The triplet $U = (U_T, U_M, U_Y)$ is a random vector that accounts for all variations among individuals.

In Figure 4.6(a), the omitted factors are assumed to be arbitrarily distributed but mutually independent. In Figure 4.6(b), the dashed arcs connecting U_T and U_M (as well as U_M and U_T) encode the understanding that the factors in question may be dependent.

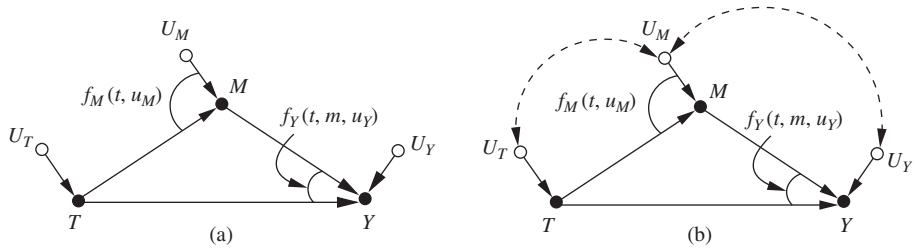

Figure 4.6 (a) The basic nonparametric mediation model, with no confounding. (b) A confounded mediation model in which dependence exists between U_M and (U_T, U_Y)

Counterfactual definition of direct and indirect effects

Using the structural model of Eq. (4.43) and the counterfactual notation defined in Section 4.2.1, four types of effects can be defined for the transition from $T = 0$ to $T = 1$. Generalizations to arbitrary reference points, say from $T = t$ to $T = t'$, are straightforward[1]:

(a) Total effect –

$$\begin{aligned} TE &= E[Y_1 - Y_0] \\ &= E[Y|do(T = 1)] - E[Y|do(T = 0)] \end{aligned} \tag{4.44}$$

TE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is allowed to track the change in T naturally, as dictated by the function f_M.

(b) Controlled direct effect –

$$\begin{aligned} CDE(m) &= E[Y_{1,m} - Y_{0,m}] \\ &= E[Y|do(T = 1, M = m)] - E[Y|do(T = 0, M = m)] \end{aligned} \tag{4.45}$$

CDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to a specified level $M = m$ uniformly over the entire population.

(c) Natural direct effect –

$$NDE = E[Y_{1,M_0} - Y_{0,M_0}] \tag{4.46}$$

NDE measures the expected increase in Y as the treatment changes from $T = 0$ to $T = 1$, while the mediator is set to whatever value it *would have attained* (for each individual) prior to the change, that is, under $T = 0$.

(d) Natural indirect effect –

$$NIE = E[Y_{0,M_1} - Y_{0,M_0}] \tag{4.47}$$

NIE measures the expected increase in Y when the treatment is held constant, at $T = 0$, and M changes to whatever value it would have attained (for each individual) under $T = 1$. It captures, therefore, the portion of the effect that can be explained by mediation alone, while disabling the capacity of Y to respond to X.

[1] These definitions apply at the population levels; the unit-level effects are given by the expressions under the expectation. All expectations are taken over the factors U_M and U_Y.

We note that, in general, the total effect can be decomposed as

$$TE = NDE - NIE_r \tag{4.48}$$

where NIE_r stands for the NIE under the reverse transition, from $T = 1$ to $T = 0$. This implies that *NIE* is identifiable whenever *NDE* and *TE* are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula, $TE = NDE + NIE$.

We further note that *TE* and $CDE(m)$ are *do*-expressions and can, therefore, be estimated from experimental data or in observational studies using the backdoor or front-door adjustments. Not so for the *NDE* and *NIE*; a new set of assumptions is needed for their identification.

Conditions for identifying natural effects

The following set of conditions, marked *A*-1 to *A*-4, are sufficient for identifying both direct and indirect natural effects.

We can identify the *NDE* and *NIE* provided that there exists a set W of measured covariates such that

A-1 No member of W is a descendant of T.
A-2 W blocks all backdoor paths from M to Y (after removing $T \rightarrow M$ and $T \rightarrow Y$).
A-3 The W-specific effect of T on M is identifiable (possibly using experiments or adjustments).
A-4 The W-specific joint effect of $\{T, M\}$ on Y is identifiable (possibly using experiments or adjustments).

Theorem 4.5.2 (Identification of the *NDE*) *When conditions A-1 and A-2 hold, the natural direct effect is experimentally identifiable and is given by*

$$NDE = \sum_m \sum_w [E[Y|do(T = 1, M = m), W = w] - E[Y|do(T = 0, M = m), W = w]] \times P(M = m|do(T = 0), W = w)P(W = w) \tag{4.49}$$

The identifiability of the do-expressions in Eq. (4.49) is guaranteed by conditions A-3 and A-4 and can be determined using the backdoor or front-door criteria.

Corollary 4.5.1 *If conditions A-1 and A-2 are satisfied by a set W that also deconfounds the relationships in A-3 and A-4, then the do-expressions in Eq. (4.49) are reducible to conditional expectations, and the natural direct effect becomes*

$$NDE = \sum_m \sum_w [E[Y|T = 1, M = m, W = w] - E[Y|T = 0, M = m, W = w]] \times P(M = m|T = 0, W = w)P(W = w) \tag{4.50}$$

In the nonconfounding case (Figure 4.6(a)), *NDE* reduces to

$$NDE = \sum_m [E[Y \mid T = 1, M = m] - E[Y \mid T = 0, M = m]]P(M = m \mid T = 0). \tag{4.51}$$

Similarly, using (4.48) and $TE = E[Y|T = 1] - E[Y|T = 0]$, *NIE* becomes

$$NIE = \sum_m E[Y \mid T = 0, M = m][P(M = m \mid T = 1) - P(M = m \mid T = 0)] \tag{4.52}$$

The last two expressions are known as the *mediation formulas*. We see that while *NDE* is a weighted average of *CDE*, no such interpretation can be given to *NIE*.

The counterfactual definitions of *NDE* and *NIE* (Eqs. (4.46) and (4.47)) permit us to give these effects meaningful interpretations in terms of "response fractions." The ratio NDE/TE measures the fraction of the response that is transmitted directly, with M "frozen." NIE/TE measures the fraction of the response that may be transmitted through M, with Y blinded to X. Consequently, the difference $(TE - NDE)/TE$ measures the fraction of the response that is necessarily due to M.

Numerical example: Mediation with binary variables

To anchor these mediation formulas in a concrete example, we return to the encouragement-design example of Section 4.2.3 and assume that $T = 1$ stands for participation in an enhanced training program, $Y = 1$ for passing the exam, and $M = 1$ for a student spending more than 3 hours per week on homework. Assume further that the data described in Tables 4.6 and 4.7 were obtained in a randomized trial with no mediator-to-outcome confounding (Figure 4.6(a)). The data shows that training tends to increase both the time spent on homework and the rate of success on the exam. Moreover, training and time spent on homework together are more likely to produce success than each factor alone.

Our research question asks for the extent to which students' homework contributes to their increased success rates regardless of the training program. The policy implications of such questions lie in evaluating policy options that either curtail or enhance homework efforts, for example, by counting homework effort in the final grade or by providing students with

Table 4.6 The expected success (Y) for treated ($T = 1$) and untreated ($T = 0$) students, as a function of their homework (M)

Treatment T	Homework M	Success rate $E(Y\|T = t, M = m)$
1	1	0.80
1	0	0.40
0	1	0.30
0	0	0.20

Table 4.7 The expected homework (M) done by treated (T =1) and untreated (T =0) students

Treatment T	Homework $E(M\|T = t)$
0	0.40
1	0.75

adequate work environments at home. An extreme explanation of the data, with significant impact on educational policy, might argue that the program does not contribute substantively to students' success, save for encouraging students to spend more time on homework, an encouragement that could be obtained through less expensive means. Opposing this theory, we may have teachers who argue that the program's success is substantive, achieved mainly due to the unique features of the curriculum covered, whereas the increase in homework efforts cannot alone account for the success observed.

Substituting the data into Eqs. (4.51) and (4.52) gives

$$NDE = (0.40 - 0.20)(1 - 0.40) + (0.80 - 0.30)0.40 = 0.32$$

$$NIE = (0.75 - 0.40)(0.30 - 0.20) = 0.035$$

$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times 0.60) = 0.46$$

$$NIE/TE = 0.07, NDE/TE = 0.696, 1 - NDE/TE = 0.304$$

We conclude that the program as a whole has increased the success rate by 46% and that a significant portion, 30.4%, of this increase is due to the capacity of the program to stimulate improved homework effort. At the same time, only 7% of the increase can be explained by stimulated homework alone without the benefit of the program itself.

Study questions

Study question 4.5.2

Consider the structural model:

$$y = \beta_1 m + \beta_2 t + u_y \tag{4.53}$$

$$m = \gamma_1 t + u_m \tag{4.54}$$

(a) Use the basic definition of the natural effects (Eqs. (4.46) and (4.47)) to determine TE, NDE, and NIE.
(b) Repeat (a) assuming that u_y is correlated with u_m.

Study question 4.5.3

Consider the structural model:

$$y = \beta_1 m + \beta_2 t + \beta_3 tm + \beta_4 w + u_y \tag{4.55}$$

$$m = \gamma_1 t + \gamma_2 w + u_m \tag{4.56}$$

$$w = \alpha t + u_w \tag{4.57}$$

with $\beta_3 tm$ representing an interaction term.

(a) Use the basic definition of the natural effects (Eqs. (4.46) and (4.47)) (treating M as the mediator), to determine the portion of the effect for which mediation is necessary *(TE − NDE) and the portion for which mediation is* sufficient *(NIE). Hint: Show that:*

$$NDE = \beta_2 + \alpha\beta_4 \tag{4.58}$$

$$NIE = \beta_1(\gamma_1 + \alpha\gamma_2) \tag{4.59}$$

$$TE = \beta_2 + (\gamma_1 + \alpha\gamma_2)(\beta_3 + \beta_1) + \alpha\beta_4 \tag{4.60}$$

$$TE - NDE = (\beta_1 + \beta_3)(\gamma_1 + \alpha\gamma_2) \tag{4.61}$$

(b) Repeat, using W as the mediator.

Study question 4.5.4

Apply the mediation formulas provided in this section to the discrimination case discussed in Section 4.4.4, and determine the extent to which ABC International practiced discrimination in their hiring criteria. Use the data in Tables 4.6 and 4.7, with $T = 1$ standing for male applicants, $M = 1$ standing for highly qualified applicants, and $Y = 1$ standing for hiring. (Find the proportion of the hiring disparity that is due to gender, and the proportion that could be explained by disparity in qualification alone.)

Ending Remarks

The analysis of mediation is perhaps the best arena to illustrate the effectiveness of the counterfactual-graphical symbiosis that we have been pursuing in this book. If we examine the identifying conditions *A*-1 to *A*-4, we find four assertions about the model that are not too easily comprehended. To judge their plausibility in any given scenario, without the graph before us, is unquestionably a formidable, superhuman task. Yet the symbiotic analysis frees investigators from the need to understand, articulate, examine, and judge the plausibility of the assumptions needed for identification. Instead, the method can confirm or disconfirm these assumptions algorithmically from a more reliable set of assumption, those encoded in the structural model itself. Once constructed, the causal diagram allows simple path-tracing routines to replace much of the human judgment deemed necessary in mediation analysis; the judgment invoked in the construction of the diagrams is sufficient, and that construction requires only judgment about causal relationships among realizable variables and their disturbances.

Bibliographical Notes for Chapter 4

The definition of counterfactuals as derivatives of structural equations, Eq. (4.5), was introduced by Balke and Pearl (1994a,b), who applied it to the estimation of probabilities of causation in legal settings. The philosopher David Lewis defined counterfactuals in terms of similarity among possible worlds Lewis (1973). In statistics, the notation $Y_x(u)$ was devised by Neyman (1923), to denote the potential response of unit u in a controlled randomized trial,

under treatment $X = x$. It remained relatively unnoticed until Rubin (1974) treated Y_x as a random variable and connected it to observed variable via the consistency rule of Eq. (4.6), which is a theorem in both Lewis's logic and in structural models. The relationships among these three formalisms of counterfactuals are discussed at length in Pearl (2000, Chapter 7), where they are shown to be logically equivalent; a problem solved in one framework would yield the same solution in another. Rubin's framework, known as "potential outcomes," differs from the structural account only in the language in which problems are defined, hence, in the mathematical tools available for their solution. In the potential outcome framework, problems are defined algebraically as assumptions about counterfactual independencies, also known as "ignorability assumptions." These types of assumptions, exemplified in Eq. (4.15), may become too complicated to interpret or verify by unaided judgment. In the structural framework, on the other hand, problems are defined in the form of causal graphs, from which dependencies of counterfactuals (e.g., Eq. (4.15)) can be derived mechanically. The reason some statisticians prefer the algebraic approach is, primarily, because graphs are relatively new to statistics. Recent books in social science (e.g., Morgan and Winship 2014) and in health science (e.g., VanderWeele 2015) are taking the hybrid, graph-counterfactual approach pursued in our book.

The section on linear counterfactuals is based on Pearl (2009, pp. 389–391). Recent advances are provided in Cai and Kuroki (2006) and Chen and Pearl (2014). Our discussion of ETT (Effect of Treatment on the Treated), as well as additive interventions, is based on Shpitser and Pearl (2009), which provides a full characterization of models in which ETT is identifiable.

Legal questions of attribution, as well as probabilities of causation are discussed at length in Greenland (1999) who pioneered the counterfactual approach to such questions. Our treatment of PN, PS, and PNS is based on Tian and Pearl (2000) and Pearl (2000, Chapter 9). Recent results, including the tool kit of Section 4.5.1, are given in Pearl (2015a).

Mediation analysis (Sections 4.4.5 and 4.5.2), as we remarked in Chapter 3, has a long tradition in the social sciences (Duncan 1975; Kenny 1979), but has gone through a dramatic revolution through the introduction of counterfactual analysis. A historical account of the conceptual transition from the statistical approach of Baron and Kenny (1986) to the modern, counterfactual-based approach of natural direct and indirect effects (Pearl 2001; Robins and Greenland 1992) is given in Sections 1 and 2 of Pearl (2014a). The recent text of VanderWeele (2015) enhances this development with new results and new applications. Additional advances in mediation, including sensitivity analysis, bounds, multiple mediators, and stronger identifying assumptions are discussed in Imai et al. (2010) and Muthén and Asparouhov (2015).

The mediation tool kit of Section 4.5.2 is based on Pearl (2014a). Shpitser (2013) has derived a general criterion for identifying indirect effects in graphs.