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.

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

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

and we sought to estimate them in terms of observed probabilities such as $P(X = x, Y = y)$. 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 $E[Y_1] = E[Y|do(\textit{freeway})]$ and $E[Y_0] = E[Y|do(\textit{Sepulveda})]$, 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 a 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 2 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 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

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

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,