

Consider for example the problem of finding the best estimate of Z given two observations, $X = x$ and $Y = y$. As before, we write the regression equation

$$Z = \alpha + \beta_Y Y + \beta_X X + \epsilon$$

But now, to obtain three equations for α , β_Y , and β_X , we also multiply both sides by Y and X and take expectations. Imposing the orthogonality conditions $E[\epsilon Y] = E[\epsilon X] = 0$ and solving the resulting equations gives

$$\beta_Y = R_{ZY \cdot X} = \frac{\sigma_X^2 \sigma_{ZY} - \sigma_{ZX} \sigma_{XY}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \quad (1.27)$$

$$\beta_X = R_{ZX \cdot Y} = \frac{\sigma_Y^2 \sigma_{ZX} - \sigma_{ZY} \sigma_{YX}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \quad (1.28)$$

Equations (1.27) and (1.28) are generic; they give the linear regression coefficients $R_{ZY \cdot X}$ and $R_{ZX \cdot Y}$ for any three variables in terms of their variances and covariances, and as such, they allow us to see how sensitive these slopes are to other model parameters. In practice, however, regression slopes are estimated from sampled data by efficient “least-square” algorithms, and rarely require memorization of mathematical equations. An exception is the task of predicting whether any of these slopes is zero, prior to obtaining any data. Such predictions are important when we contemplate choosing a set of regressors for one purpose or another, and as we shall see in Section 3.8, this task will be handled quite efficiently through the use of causal graphs.

Study question 1.3.9

- (a) Prove Eq. (1.22) using the orthogonality principle. [Hint: Follow the treatment of Eq. (1.26).]
- (b) Find all partial regression coefficients

$$R_{YX \cdot Z}, R_{XY \cdot Z}, R_{YZ \cdot X}, R_{ZY \cdot X}, R_{XZ \cdot Y}, \text{ and } R_{ZX \cdot Y}$$

for the craps game described in Study question 1.3.7. [Hint: Apply Eq. (1.27) and use the variances and covariances computed for part (a) of this question.] **Study question 1.3.8.**

1.4 Graphs

We learned from Simpson’s Paradox that certain decisions cannot be made on the basis of data alone, but instead depend on the story behind the data. In this section, we layout a mathematical language, *graph theory*, in which these stories can be conveyed. Graph theory is not generally taught in high school mathematics, but it provides a useful mathematical language that allows us to address problems of causality with simple operations similar to those used to solve arithmetic problems.

Although the word *graph* is used colloquially to refer to a whole range of visual aids—more or less interchangeably with the word *chart*—in mathematics, a graph is a formally defined

If we use our product decomposition rule, however, the 45 samples are separated into much larger categories. In order to determine $P(x)$, every (x, y, z, w) sample falls into one of only two cells: $(X = 1)$ and $(X = 0)$. Clearly, the probability of leaving either of them empty is much lower, and the accuracy of estimating population frequencies is much higher. The same is true of the divisions we need to make to determine $P(y|x) : (Y = 1, X = 1), (Y = 0, X = 1), (Y = 1, X = 0),$ and $(Y = 0, X = 0)$. And to determine $P(z|y) : (Y = 1, Z = 1), (Y = 0, Z = 1), (Y = 1, Z = 0),$ and $(Y = 0, Z = 0)$. And to determine $P(w|z) : (W = 1, Z = 1), (W = 0, Z = 1), (W = 1, Z = 0),$ and $(W = 0, Z = 0)$. Each of these divisions will give us much more accurate frequencies than our original division into 15 cells. Here we explicitly see the simpler estimation problems allowed by assuming the graphical structure of an SCM and the resulting improved accuracy of our frequency estimates.

This is not the only use to which we can put the qualitative knowledge that a graph provides. As we will see in the next section, graphical models reveal much more information than is obvious at first glance; we can learn a lot about, and infer a lot from, a data set using only the graphical model of its causal story.

Study questions

Study question 1.5.2

Assume that a population of patients contains a fraction r of individuals who suffer from a certain fatal syndrome Z , which simultaneously makes it uncomfortable for them to take a life-prolonging drug X (Figure 1.10). Let $Z = z_1$ and $Z = z_0$ represent, respectively, the presence and absence of the syndrome, $Y = y_1$ and $Y = y_0$ represent death and survival, respectively, and $X = x_1$ and $X = x_0$ represent taking and not taking the drug. Assume that patients not carrying the syndrome, $Z = z_0$, die with probability p_2 if they take the drug and with probability p_1 if they don't. Patients carrying the syndrome, $Z = z_1$, on the other hand, die with probability p_3 if they do not take the drug and with probability p_4 if they do take the drug. Further, patients having the syndrome are more likely to avoid the drug, with probabilities $q_1 = P(x_1|z_0)$ and $q_2 = P(x_0|z_1)$.

- (a) Based on this model, compute the joint distributions $P(x, y, z)$, $P(x, y)$, $P(x, z)$, and $P(y, z)$ for all values of x, y , and z , in terms of the parameters $(r, p_1, p_2, p_3, p_4, q_1, q_2)$. [Hint: Use the product decomposition of Section 1.5.2.]
- (b) Calculate the difference $P(y_1|x_1) - P(y_1|x_0)$ for three populations: (1) those carrying the syndrome, (2) those not carrying the syndrome, and (3) the population as a whole.

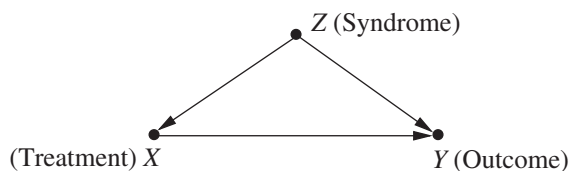


Figure 1.10 Model showing an unobserved syndrome, Z , affecting both treatment (X) and outcome (Y)

its value; there is some case in which changing the value of the first variable changes the value of the second. That means that when we examine those variables in the data set, the probability that one variable takes a given value will change, given that we know the value of the other variable. So in any causal model, regardless of the specific functions, two variables connected by an edge are dependent. By this reasoning, we can see that in SCMs 2.2.1–2.2.3, Z and Y are dependent, and Y and X are dependent.

From these two facts, we can conclude that Z and X are *likely* dependent. If Z depends on Y for its value, and Y depends on X for its value, then Z likely depends on X for its value. There are pathological cases in which this is not true. Consider, for example, the following SCM, which also has the graph in Figure 2.1.

SCM 2.2.4 (Pathological Case of Intransitive Dependence)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} a & \text{IF } X = 1 \text{ AND } U_Y = 1 \\ b & \text{IF } X = 2 \text{ AND } U_Y = 1 \\ c & \text{IF } U_Y = 2 \end{cases}$$

$$f_Z : Z = \begin{cases} i & \text{IF } Y = c \text{ OR } U_Z = 1 \\ j & \text{IF } U_Z = 2 \end{cases}$$

In this case, no matter what value U_Y and U_Z take, X will have no effect on the value that Z takes; changes in X account for variation in Y between a and b , but Y doesn't affect Z unless it takes the value c . Therefore, X and Z vary independently in this model. We will call cases such as these *intransitive cases*.

However, intransitive cases form only a small number of the cases we will encounter. In most cases, the values of X and Z vary together just as X and Y do, and Y and Z . Therefore, they are likely dependent in the data set.

Now, let's consider point 4: Z and X are independent conditional on Y . Remember that when we condition on Y , we filter the data into groups based on the value of Y . So we compare all the cases where $Y = a$, all the cases where $Y = b$, and so on. Let's assume that we're looking at the cases where $Y = a$. We want to know whether, *in these cases only*, the value of Z is independent of the value of X . Previously, we determined that X and Z are likely dependent, because when the value of X changes, the value of Y likely changes, and when the value of Y changes, the value of Z is likely to change. Now, however, examining *only the cases where* $Y = a$, when we select cases with different values of X , the value of U_Y changes so as to keep Y at $Y = a$, but since Z depends only on Y and U_Z , not on U_Y , the value of Z remains unaltered. So selecting a different value of X doesn't change the value of Z . So, in the case where $Y = a$, X is independent of Z . This is of course true no matter which specific value of Y we condition on. So X is independent of Z , conditional on Y .

This configuration of variables—three nodes and two edges, with one edge directed into and one edge directed out of the middle variable—is called a *chain*. Analogous reasoning to the above tells us that in any graphical model, given any two variables X and Y , if the only path between X and Y is composed entirely of chains, then X and Y are independent conditional on any intermediate variable on that path. This independence relation holds regardless of the functions that connect the variables. This gives us a rule:

This occurs for example when X and U_Y are fair coins and $Y = 1$ if and only $X = U_Y$. In this case $P(Y=1|X=1) = P(Y=1|X=0) = P(Y=1) = 1/2$. Such pathological cases require precise numerical probabilities to achieve independence ($P(X=1) = P(U_Y) = 1/2$); they are rare, and can be ignored for all practical purposes.

directly from the data set, using the procedures described in Part One, while the former varies depending on the structure of the causal graph. It is the graph that instructs us which arrow should be removed for any given intervention.

In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $P(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x . On the other hand, $P(Y = y|do(X = x))$ represents the population distribution of Y if *everyone in the population* had their X value fixed at x . We similarly write $P(Y = y|do(X = x), Z = z)$ to denote the conditional probability of $Y = y$, given $Z = z$, in the distribution created by the intervention $do(X = x)$.

Using *do*-expressions and graph surgery, we can begin to untangle the causal relationships from the correlative. In the rest of this chapter, we learn methods that can, astoundingly, tease out causal information from purely observational data, assuming of course that the graph constitutes a valid representation of reality. It is worth noting here that we are making a tacit assumption ~~here~~ that the intervention has no “side effects,” that is, that *assigning* the value x for the ~~variable~~ ~~variable~~ X for an individual does not alter subsequent variables in a direct way. For example, being “assigned” a drug might have a different effect on recovery than being forced to take the drug against one’s religious objections. When side effects are present, they need to be specified explicitly in the model.

3.2 The Adjustment Formula

The ice cream example represents an extreme case in which the correlation between X and Y was totally spurious from a causal perspective, because there was no causal path from X to Y . Most real-life situations are not so clear-cut. To explore a more realistic situation, let us examine Figure 3.3, in which Y responds to both Z and X . Such a model could represent, for example, the first story we encountered for Simpson’s paradox, where X stands for drug usage, Y stands for recovery, and Z stands for gender. To find out how effective the drug is in the population, we imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what would obtain under the complementary intervention, where we prevent everyone from using the drug. Denoting the first intervention by $do(X = 1)$ and the second by $do(X = 0)$, our task is to estimate the difference

$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{3.1}$$

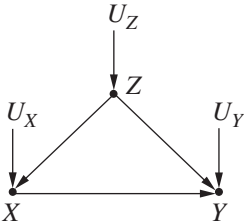


Figure 3.3 A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

$T \rightarrow Y$. This path is spurious since it lies outside the causal pathway from X to Y . Opening this path will create bias and yield an erroneous answer. This means that computing the association between X and Y for each value of W separately will not yield the correct effect of X on Y , and it might even give the wrong effect for each value of W .

How then do we compute the causal effect of X on Y for a specific value w of W ? W may represent, for example, the level of posttreatment pain of a patient, and we might be interested in assessing the effect of X on Y for only those patients who did not suffer any pain. Specifying the value of W amounts to conditioning on $W = w$, and this, as we have realized, opens a spurious path from X to Y by virtue of the fact that W is a collider.

The answer is that we still have the option of blocking that path using other variables. For example, if we condition on T , we would block the spurious path $X \rightarrow W \leftarrow Z \leftarrow T \rightarrow Y$, even if W is part of the conditioning set. Thus to compute the w -specific causal effect, written $P(y|do(x), w)$, we adjust for T , and obtain

$$P(Y = y|do(X = x), W = w) = \sum_t P(Y = y|X = x, W = w, T = t)P(T = t|W = w) \quad (3.11)$$

Computing such W -specific causal effects is an essential step in examining *effect modification* or *moderation*, that is, the degree to which the causal effect of X on Y is modified by different values of W . Consider, again, the model in Figure 3.6, and suppose we wish to test whether the causal effect for units at level $W = w$ is the same as for units at level $W = w'$ (W may represent any pretreatment variable, such as age, sex, or ethnicity). This question calls for comparing two causal effects,

$$P(Y = y|do(X = x), W = w) \quad \text{and} \quad P(Y = y|do(X = x), W = w')$$

In the specific example of Figure 3.6, the answer is simple, because W satisfies the backdoor criterion. So, all we need to compare are the conditional probabilities $P(Y = y|X = x, W = w)$ and $P(Y = y|X = x, W = w')$; no summation is required. In the more general case, where W alone does not satisfy the backdoor criterion, yet a larger set, $T \cup W$, does, we need to adjust for members of T , which yields Eq. (3.11). We will return to this topic in Section 3.5.

From the examples seen thus far, readers may get the impression that one should refrain from adjusting for colliders. Such adjustment is sometimes unavoidable, as seen in Figure 3.7. Here, there are four backdoor paths from X to Y , all traversing variable Z , which is a collider on the path $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$. Conditioning on Z will unblock this path and will violate the backdoor criterion. To block all backdoor paths, we need to condition on one of the following sets: $\{E, Z\}$, $\{A, Z\}$, or $\{E, Z, A\}$. Each of these contains Z . We see, therefore, that Z , a collider, must be adjusted for in any set that yields an unbiased estimate of the effect of X on Y .

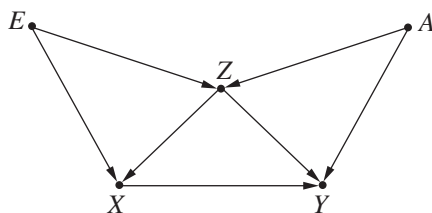


Figure 3.7 A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y



Assume that a careful study was undertaken, in which the following factors were measured simultaneously on a randomly selected sample of 800,000 subjects considered to be at very high risk of cancer (because of environmental exposures such as smoking, asbestos, radon, and the like).

- 1. Whether the subject smoked
- 2. Amount of tar in the subject's lungs
- 3. Whether lung cancer has been detected in the patient.

The data from this study are presented in Table 3.1, where, for simplicity, all three variables are assumed to be binary. All numbers are given in thousands.

Table 3.1 A hypothetical data set of randomly selected samples showing the percentage of cancer cases for smokers and nonsmokers in each tar category (numbers in thousands)

	Tar 400		No tar 400		All subjects 800	
	Smokers	Nonsmokers	Smokers	Nonsmokers	Smokers	Nonsmokers
No cancer	380	20	20	380	400	400
	323	1	18	38	341	39
	(85%)	(5%)	(90%)	(10%)	(85%)	(9.75%)
Cancer	57	19	2	342	59	361
	(15%)	(95%)	(10%)	(90%)	(15%)	(90.25%)

Two opposing interpretations can be offered for these data. The tobacco industry argues that the table proves the beneficial effect of smoking. They point to the fact that only 15% of the smokers have developed lung cancer, compared to 90.25% of the nonsmokers. Moreover, within each of two subgroups, tar and no tar, smokers show a much lower percentage of cancer than nonsmokers. (These numbers are obviously contrary to empirical observations but well illustrate our point that observations are not to be trusted.)

However, the antismoking lobbyists argue that the table tells an entirely different story—that smoking would actually increase, not decrease, one's risk of lung cancer. Their argument goes as follows: If you choose to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you choose not to smoke (380/400 vs 20/400). To evaluate the effect of tar deposits, we look separately at two groups, smokers and nonsmokers, as done in Table 3.2. All numbers are given in thousands.

Table 3.2 Reorganization of the data set of Table 3.1 showing the percentage of cancer cases in each smoking-tar category (numbers in thousands)

	Smokers 400		Nonsmokers 400		All subjects 800	
	Tar	No tar	Tar	No tar	Tar	No tar
No cancer	380	20	20	380	400	400
	323	18	1	38	324	56
	(85%)	(90%)	(5%)	(10%)	(81%)	(19%)
Cancer	57	2	19	342	76	344
	(15%)	(10%)	(95%)	(90%)	(19%)	(81%)

content, I could determine that a fresh bottle would offer the average patient a greater chance of recovery.

Based on two very reasonable assumptions, the data show clearly that the fresh drug is more effective. The assumptions are as follows:

- (i) Customers had no information about the chemical content (high or low) of the specific bottle of the drug that they were buying; their choices were influenced by price and shelf-age alone.
 - (ii) The effect of the drug on any given individual depends only on its chemical content, not on its shelf age (fresh or aged).
- (a) Determine the relevant variables for the problem, and describe this scenario in a causal graph.
 - (b) Construct a data set compatible with the story and the decision to buy the expensive bottle.
 - (c) Determine the effect of choosing the fresh versus the aged drug by using assumptions (i) and (ii), and the data given in (b).

3.5 Conditional Interventions and Covariate-Specific Effects

The interventions considered thus far have been limited to actions that merely force a variable or a group of variables X to take on some specified value x . In general, interventions may involve dynamic policies in which a variable X is made to respond in a specified way to some set Z of other variables—say, through a functional relationship $x = g(z)$ or through a stochastic relationship, whereby X is set to x with probability $P^*(x|z)$. For example, suppose a doctor decides to administer a drug only to patients whose temperature Z exceeds a certain level, $Z = z$. In this case, the action will be *conditional* upon the value of Z and can be written $do(X = g(Z))$, where $g(Z)$ is equal to one when $Z > z$ and zero otherwise (where $X = 0$ represents no drug). Since Z is a random variable, the value of X chosen by the action will similarly be a random variable, tracking variations in Z . The result of implementing such a policy is a probability distribution written $P(Y = y|do(X = g(Z)))$, which depends only on the function g and the set Z of variables that drive X .

In order to estimate the effect of such a policy, let us take a closer look at another concept, the “ z -specific effect” of X , which we encountered briefly in Section 3.3 (Eq. (3.11)). This effect, written $P(Y = y|do(X = x), Z = z)$, measures the distribution of Y in a subset of the population for which Z achieves the value z after the intervention. For example, we may be interested in how a treatment affects a specific age group, $Z = z$, or people with a specific feature, $Z = z$, which may be measured after the treatment.

The z -specific effect can be identified by a procedure similar to the backdoor adjustment. The reasoning goes as follows: When we aim to estimate $P(Y = y|do(X = x))$, an adjustment for a set S is justified if S blocks all backdoor paths from X to Y . Now that we wish to identify $P(Y = y|do(X = x), Z = z)$, we need to ensure that those paths remain blocked when we add one more variable, Z , to the conditioning set. This yields a simple criterion for the identification of the z -specific effect:

Rule 2 The z -specific effect $P(Y = y|do(X = x), Z = z)$ is identified whenever we can measure a set S of variables such that $S \cup Z$ satisfies the backdoor criterion. Moreover, the z -specific

effect is given by the following adjustment formula

$$P(Y = y | do(X = x), Z = z) = \sum_s P(Y = y | X = x, S = s, Z = z) P(S = s \mid Z = z)$$

This modified adjustment formula is similar to Eq. (3.5) with two exceptions. First, the adjustment set is $S \cup Z$, not just S and, second, the summation goes only over S , not including Z . The $\cup$ symbol in the expression $S \cup Z$ stands for set addition (or union), which means that, if Z is a subset of S , we have $S \cup Z = S$, and S alone need satisfy the backdoor criterion.

Note that the identifiability criterion for z -specific effects is somewhat stricter than that for nonspecific effect. Adding Z to the conditioning set might create dependencies that would prevent the blocking of all backdoor paths. A simple example occurs when Z is a collider; conditioning on Z will create a new dependency between Z 's parents and thus violate the backdoor requirement.

We are now ready to tackle our original task of estimating conditional interventions. Suppose a policy maker contemplates an age-dependent policy whereby an amount x of drug is to be administered to patients, depending on their age Z . We write it as $do(X = g(Z))$. To find out the distribution of outcome Y that results from this policy, we seek to estimate $P(Y = y | do(X = g(Z)))$.

We now show that identifying the effect of such policies is equivalent to identifying the expression for the z -specific effect $P(Y = y | do(X = x), Z = z)$.

To compute $P(Y = y | do(X = g(Z)))$, we condition on $Z = z$ and write

$$\begin{aligned} P(Y = y | do(X = g(Z))) &= \sum_z P(Y = y | do(X = g(Z)), Z = z) P(Z = z | do(X = g(Z))) \\ &= \sum_z P(Y = y | do(X = g(z)), Z = z) P(Z = z) \end{aligned} \quad (3.17)$$

The equality

$$P(Z = z | do(X = g(Z))) = P(Z = z)$$

stems, of course, from the fact that Z occurs before X ; hence, any control exerted on X can have no effect on the distribution of Z . Equation (3.17) can also be written as

$$\sum_z P(Y = y | do(X = x), \overset{Z=z}{\cancel{Z}}) |_{x=g(z)} P(Z = z)$$

which tells us that the causal effect of a conditional policy $do(X = g(Z))$ can be evaluated directly from the expression of $P(Y = y | do(X = x), Z = z)$ simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(Z = z)$).

Study question 3.5.1

Consider the causal model of Figure 3.8.

(a) Find an expression for the c -specific effect of X on Y .

from Bayes’ rule, which tells us that

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

In other words, to find the probability of each row in the surviving table, we multiply the unconditional probability, $P(Y = y, Z = z, X = x)$ by the constant $1/P(X = x)$.

Let us now examine the population created by the $do(X = x)$ operation and ask how the probability of each case changes as a result of this operation. The answer is given to us by the adjustment formula, which reads

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

Multiplying and dividing the expression inside the sum by the propensity score $P(X = x|Z = z)$, we get



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

Upon realizing the numerator is none other but the pretreatment distribution of (X, Y, Z) , we can write

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



and the answer becomes clear: each case $(Y = y, X = x, Z = z)$ in the population should boost its probability by a factor equals to $1/P(X = x|Z = z)$. (Hence the name “inverse probability ~~weighting~~ ^{weighting}.”)

This provides us with a simple procedure of estimating $P(Y = y|do(X = x))$ when we have finite samples. If we weigh each available sample by a factor $= 1/P(X = x|Z = z)$, we can then treat the reweighted samples as if they were generated from P_m , not P , and proceed to estimate $P(Y = y|do(x))$ accordingly.

This is best demonstrated in an example.

Table 3.3 returns to our Simpson’s paradox example of the drug that seems to help men and women but to hurt the general population. We’ll use the same data we used before but presented

Table 3.3 Joint probability distribution $P(X, Y, Z)$ for the drug-gender-recovery story of Chapter 1 (Table 1.1)

X	Y	Z	% of population
Yes	Yes	Male	0.116
Yes	Yes	Female	0.274
Yes	No	Male	0.01 0.009
Yes	No	Female	0.101
No	Yes	Male	0.334
No	Yes	Female	0.079
No	No	Male	0.051
No	No	Female	0.036

Table 3.4 Conditional probability distribution $P(Y, Z|X)$ for drug users ($X = \text{yes}$) in the population of Table 3.3

X	Y	Z	% of population
Yes	Yes	Male	0.232 $\wedge$ 0.231
Yes	Yes	Female	0.547 $\wedge$ 0.549
Yes	No	Male	0.02 $\wedge$ 0.017
Yes	No	Female	0.202 $\wedge$ 0.203

this time as a weighted table. In this case, X represents whether or not the patient took the drug, Y represents whether the patient recovered, and Z represents the patient's gender.

If we condition on " $X = \text{yes}$," we get the data set shown in Table 3.4, which was formed in two steps. First, all rows with $X = \text{No}$ were excluded. Second, the weights given to the remaining rows were "renormalized," that is, multiplied by a constant so as to make them sum to one. This constant, according to Bayes' rule, is $1/P(X = \text{yes})$, and $P(X = \text{yes})$ in our example, is the combined weight of the first four rows of Table 3.3, which amounts to

$$P(X = \text{yes}) = 0.116 + 0.274 + 0.01 + 0.101 = \underset{\wedge}{0.501}$$

The result is the weight distribution in the four ~~top~~ rows of Table 3.4; the weight of each row has been boosted by a factor $\underset{\wedge}{1/0.501 = 2.00}$.

Let us now examine the population created by the $do(X = \text{yes})$ operation, representing a deliberate decision to administer the drug to the same population.

To calculate the distribution of weights in this population, we need to compute the factor $P(X = \text{yes}|Z = z)$ for each z , which, according to Table 3.3, is given by

$$P(X = \text{yes}|Z = \text{Male}) = \frac{(0.116 + 0.01)}{(0.116 + 0.01 + 0.334 + 0.051)} = \underset{\wedge}{0.247}$$

$$P(X = \text{yes}|Z = \text{Female}) = \frac{(0.274 + 0.101)}{(0.274 + 0.101 + 0.079 + 0.036)} = 0.765$$

Multiplying the gender-matching rows by $1/\underset{\wedge}{0.247}$ and $1/0.765$, respectively, we obtain Table 3.5, which represents the postintervention distribution of the population of Table 3.3. The probability of recovery in this distribution can now be computed directly from the data, by summing the first two rows:

$$P(Y = \text{yes}|do(X = \text{yes})) = 0.476 + 0.357 = 0.833$$

Table 3.5 Probability distribution for the population of Table 3.3 under the intervention $do(X = \text{Yes})$, determined via the inverse probability method

X	Y	Z	% of population
Yes	Yes	Male	0.476 $\wedge$ 0.475
Yes	Yes	Female	0.357 $\wedge$ 0.358
Yes	No	Male	0.041 $\wedge$ 0.035
Yes	No	Female	0.132 $\wedge$ 0.132

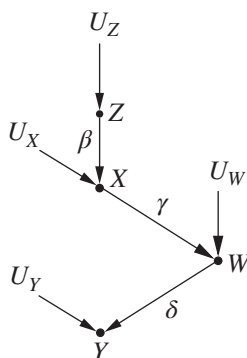


Figure 3.16 By removing the direct edge from X to Y and finding the set of variables $\{W\}$ that d -separate them, we find the variables we need to adjust for to determine the direct effect of X on Y

➤ ^{other} ~~others~~, and all variables in the graph are measured, every structural parameter can be identified in this manner, namely, there is at least one identifying regression equation in which one of the coefficients corresponds to the parameter we seek to estimate. One such equation is obviously the structural equation itself, with the parents of Y serving as regressors. But there may be several other identifying equations, with possibly better features for estimation and graphical analysis can reveal them all (see Study question 3.8.1(c)). Moreover, when some variables are not measured, or when some error terms are correlated, the task of finding an identifying regression from the structural equations themselves would normally be insurmountable; the G_α procedure then becomes indispensable (see Study question 3.8.1(d)).

Remarkably, the regression rule procedure has eluded investigators for almost a century, possibly because it is extremely difficult to articulate in algebraic, nongraphical terms.

Suppose, however, there is no set of variables that d -separates X and Y in G_α . For instance, in Figure 3.17, X and Y have an unobserved common cause represented by the dashed

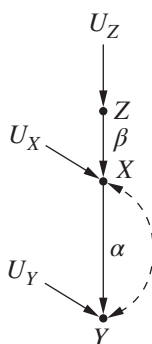


Figure 3.17 A graphical model in which we cannot find the direct effect of X on Y via adjustment, because the dashed double-arrow arc represents the presence of a backdoor path between X and Y , consisting of unmeasured variables. In this case, Z is an instrument with regard to the effect of X on Y that enables the identification of α

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

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

$$E[Y_0|Z = 1] = b \quad (4.10)$$

$$E[Y|do(X = 1), Z = 1] = b \quad \text{(see footnote 2)} \quad (4.11)$$

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

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)

		$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$						
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

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

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

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.

- 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 $E[Y_x|Z = z] = abx + bz/(1 + a^2)$.]
- 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.

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

$$Y = bX + cH + \delta XH + U_Y \quad (4.19)$$

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} of Model 2 (Figure 4.1) can be estimated from non-experimental 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

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 $add(q)$:

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

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 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:

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



close
gap

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?”

- Formulate Joe’s question mathematically, in terms of ETT, assuming that the outcome of interest is lung cancer.*
- 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.*
- 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?

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

➤
$$PNS = \overset{0.86}{\cancel{0.39}} - \overset{0.61}{\cancel{0.44}} = 0.25$$

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 ~~Sex~~ 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

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

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

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

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

$$\begin{aligned} PNS &= P(Y_x = 1, Y_{x'} = 0) \\ &= P(Y_x = 1) - P(Y_{x'} = 1) \end{aligned} \quad (4.42)$$

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 therapy) 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:

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

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.



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

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

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 <i>T</i>	Homework <i>M</i>	Success rate <i>E(Y T = t, M = m)</i>
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 <i>T</i>	Homework <i>E(M T = t)</i>
0	0.40
1	0.75

