About This Errata

This file contains changes made to the text of *Causal Inference in Statistics: A Primer* by Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Changes are pointed to with arrows in the left margins which will make it easy for you to mark your own personal copy.

If you should discover additional corrections or needed clarification, please let us know (kaoru@cs.ucla.edu).

Most of these errors were discovered by:
Andrew Forney, Michael Lewis, and Adamo Vincenzo
Many thanks.

Last revised: 4.8.2020

$\longrightarrow$ Errata implemented.
$\longrightarrow$ Errata submitted to Wiley, September/October 2016. Implemented in Revised Reprint
$\longrightarrow$ Errata submitted to Wiley, October 2018. Implemented in Revised Reprint
$\longrightarrow$ Errata pending implementation by Wiley (updated Sept 2019). Implemented in Revised Reprint
$\longrightarrow$ New errata (since January 2020)

His groundbreaking book, *Causality: Models, Reasoning and Inference* (Cambridge, 2000, 2009), has introduced many of the methods used in modern causal analysis. It won the Lakatos Award from the London School of Economics and is cited by more than 15,000 scientific publications. His latest book, *The Book of Why: The New Science of Cause and Effect* (with Dana Mackenzie, Basic Books, 2018), describes the impacts of causal inference to the general public.

About the Authors

Judea Pearl is Professor of Computer Science and Statistics at the University of California, Los Angeles, where he directs the Cognitive Systems Laboratory and conducts research in artificial intelligence, causal inference and philosophy of science. He is a Co-Founder and Editor of the *Journal of Causal Inference* and the author of three landmark books in inference-related areas. ~~His latest book, *Causality: Models, Reasoning and Inference* (Cambridge, 2000, 2009), has introduced many of the methods used in modern causal analysis. It won the Lakatosh Award from the London School of Economics and is cited by more than 13,000 scientific publications.~~

new 2nd paragraph

Pearl is a member of the National Academy of Sciences, the National Academy of Engineering, and a Founding Fellow of the Association for Artificial Intelligence. He is a recipient of numerous prizes and awards, including the Technion's Harvey Prize and the ACM Alan Turing Award for fundamental contributions to probabilistic and causal reasoning.

Madelyn Glymour is a data analyst at Carnegie Mellon University, and a science writer and editor for the Cognitive Systems Laboratory at UCLA. Her interests lie in causal discovery and in the art of making complex concepts accessible to broad audiences.

Nicholas P. Jewell is Professor of Biostatistics and Statistics at the University of California, Berkeley. He has held various academic and administrative positions at Berkeley since his arrival in 1981, most notably serving as Vice Provost from 1994 to 2000. He has also held academic appointments at the University of Edinburgh, Oxford University, the London School of Hygiene and Tropical Medicine, and at the University of Kyoto. In 2007, he was a Fellow at the Rockefeller Foundation Bellagio Study Center in Italy.

Jewell is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science (AAAS). He is a past winner of the Snedecor Award and the Marvin Zelen Leadership Award in Statistical Science from Harvard University. He is currently the Editor of the *Journal of the American Statistical Association – Theory & Methods*, and Chair of the Statistics Section of AAAS. His research focuses on the application of statistical methods to infectious and chronic disease epidemiology, the assessment of drug safety, time-to-event analyses, and human rights.

The book is organized in four chapters.

Chapter 1 provides the basic statistical, probabilistic, and graphical concepts that readers will need to understand the rest of the book. It also introduces the fundamental concepts of causality, including the causal model, and explains through examples how the model can convey information that pure data are unable to provide.

Chapter 2 explains how causal models are reflected in data, through patterns of statistical dependencies. It explains how to determine whether a data set complies with a given causal model, and briefly discusses how one might search for models that explain a given data set.

Chapter 3 is concerned with how to make predictions using causal models, with a particular emphasis on predicting the outcome of a policy intervention. Here we introduce techniques of reducing confounding bias using adjustment for covariates, as well as inverse probability weighing. This chapter also covers mediation analysis and contains an in-depth look at how the causal methods discussed thus far work in a linear system. Key to these methods is the fundamental distinction between regression coefficients and structural parameters, and how students should use both to predict causal effects in linear models.

Chapter 4 introduces the concept of counterfactuals—what would have happened, had we chosen differently at a point in the past—and discusses how we can compute them, estimate their probabilities, and what practical questions we can answer using them. This chapter is somewhat advanced, compared to its predecessors, primarily due to the novelty of the notation and the hypothetical nature of the questions asked. However, the fact that we read and compute counterfactuals using the same scientific models that we used in previous chapters should make their analysis an easy journey for students and instructors. Those wishing to understand counterfactuals on a friendly mathematical level should find this chapter a good starting point, and a solid basis for bridging the model-based approach taken in this book with the potential outcome framework that some experimentalists are pursuing in statistics.

Acknowledgments

This book is an outgrowth of a graduate course on causal inference that the first author has been teaching at UCLA in the past 20 years. It owes many of its tools and examples to former members of the Cognitive Systems Laboratory who participated in the development of this material, both as researchers and as teaching assistants. These include Alex Balke, David Chickering, David Galles, Dan Geiger, Moises Goldszmidt, Jin Kim, George Rebane, Ilya Shpitser, Jin Tian, and Thomas Verma.

We are indebted to many colleagues from whom we have learned much about causal problems, their solutions, and how to present them to general audiences. These include Clark and Maria Glymour, for providing patient ears and sound advice on matters of both causation and writing, Felix Elwert and Tyler VanderWeele for insightful comments on an earlier version of the manuscript, and the many visitors and discussants to the UCLA Causality blog who kept the discussion lively, occasionally controversial, but never boring (causality.cs.ucla.edu/blog).

Elias Bareinboim, Bryant Chen, Andrew Forney, Ang Li, Karthika Mohan, reviewed the text for accuracy and transparency. Ang and Andrew also wrote solutions to the study questions, ~~which will be available on the book's website~~ which are available to instructors from the publisher, see <http://bayes.cs.ucla.edu/PRIMER/CIS-Manual-PUBLIC.pdf>.

The answer is nowhere to be found in simple statistics. In order to decide whether the drug will harm or help a patient, we first have to understand the story behind the data—the causal mechanism that led to, or *generated*, the results we see. For instance, suppose we knew an additional fact: Estrogen has a negative effect on recovery, so women are less likely to recover than men, regardless of the drug. In addition, as we can see from the data, women are significantly *more* likely to take the drug than men are. So, the reason the drug appears to be harmful overall is that, if we select a drug user at random, that person is more likely to be a woman and hence less likely to recover than a random person who does not take the drug. Put differently, being a woman is a common cause of both drug taking and failure to recover. Therefore, to assess the effectiveness, we need to compare subjects of the same gender, thereby ensuring that any difference in recovery rates between those who take the drug and those who do not is not ascribable to estrogen. This means we should consult the segregated data, which shows us unequivocally that the drug is helpful. This matches our intuition, which tells us that the segregated data is "more specific," hence more informative, than the unsegregated data.

With a few tweaks, we can see how the same reversal can occur in a continuous example. Consider a study that measures weekly exercise and cholesterol in various age groups. When we plot exercise on the *X*-axis and cholesterol on the *Y*-axis and segregate by age, as in Figure 1.1, we see that there is a general trend downward in each group; the more young people exercise, the lower their cholesterol is, and the same applies for middle-aged people and the elderly. If, however, we use the same scatter plot, but we don't segregate by age (as in Figure 1.2), we see a general trend upward; the more a person exercises, the higher their cholesterol is. To resolve this problem, we once again turn to the story behind the data. If we know that older people, who are more likely to exercise (Figure 1.1), are also more likely to have high cholesterol regardless of exercise, then the reversal is easily explained, and easily resolved. Age is a common cause of both treatment (exercise) and outcome (cholesterol). So we should look at the age-segregated data in order to compare same-age people and thereby eliminate the possibility that the high exercisers in each group we examine are more likely to have high cholesterol due to their age, and not due to exercising.

However, and this might come as a surprise to some readers, segregated data does not always give the correct answer. Suppose we looked at the same numbers from our first example of drug taking and recovery, instead of recording participants' gender, patients' blood pressure were
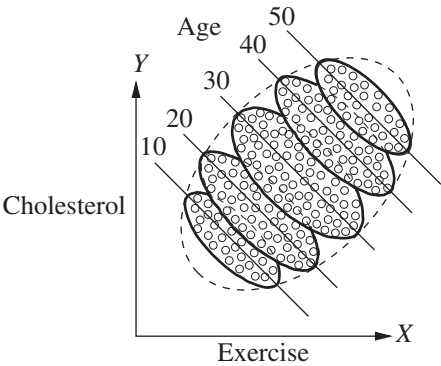

Figure 1.1   Results of the exercise–cholesterol study, segregated by age

Table 1.3   Age breakdown of voters in 2012 election (all numbers in thousands)

Age group	# of voters
18–29	20,539
30–44	30,756
45–64	52,013
65+	29,641
	132,948

Table 1.4   Age breakdown of voters over the age of 29 in 2012 election (all numbers in thousands)

Age group	# of voters
30–44	30,756
45–64	52,013
65+	29,641
	112,409

the data to form a new set (shown in Table 1.4), using only the cases where voters were older than 29.

In this new data set, there are 112,409,000 total votes, so we would estimate that

$$P(\textit{Voter Age} < 45 | \textit{Voter Age} > 29) = \frac{30,756,000}{112,409,000} = 0.27$$

Conditional probabilities such as these play an important role in investigating causal questions, as we often want to compare how the probability (or, equivalently, risk) of an outcome changes under different filtering, or exposure, conditions. For example, how does the probability of developing lung cancer for smokers compare to the analogous probability for nonsmokers?

Study questions

Study question 1.3.2

Consider Table 1.5 showing the relationship between gender and education level in the U.S. adult population.

(a) Estimate P(High School).
(b) Estimate P(High School OR Female).
(c) Estimate P(High School | Female).
(d) Estimate P(Female | High School).

on the sum of a roll of two dice. So 11 will be the sum in $\frac{2}{36} = \frac{1}{18}$ of cases: $P(\text{"11"}|\text{"craps"}) = \frac{1}{18}$. In roulette, there are 38 equally probable outcomes, so $P(\text{"11"}|\text{"roulette"}) = \frac{1}{38}$. In this situation, there are two possible hypotheses: "craps" and "roulette." Since there are an equal number of craps and roulette games, $P(\text{"craps"}) = \frac{1}{2}$, our prior belief before we hear the "11" shout. Using the law of total probability,

$$P(\text{"11"}) = P(\text{"11"}|\text{"craps"})P(\text{"craps"}) + P(\text{"11"}|\text{"roulette"})P(\text{"roulette"})$$

$$= \frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{38} = \frac{7}{171}$$

We have now fairly easily obtained all the information we need to determine $P(\text{"craps"}|\text{"11"})$:

$$P(\text{"craps"}|\text{"11"}) = \frac{P(\text{"11"}|\text{"craps"}) \times P(\text{"craps"})}{P(\text{"11"})} = \frac{1/18 \times 1/2}{7/171} = 0.679$$

Another informative example of Bayes' rule in action is the Monty Hall problem, a classic brain teaser in statistics. In the problem, you are a contestant on a game show, hosted by Monty Hall. Monty shows you three doors—A, B, and C—behind one and only one of which is a new car. (The other two doors have goats.) If you guess correctly, the car is yours; otherwise, you get a goat. You guess A at random. Monty, who is forbidden from revealing where the car is, then opens Door C, which, of course, has a goat behind it. He tells you that you can now switch to Door B, or stick with Door A. Whichever you pick, you'll get what's behind it.

Are you better off opening Door A, or switching to Door B?

Many people, when they first encounter the problem, reason that, since the location of the car is independent of the door you first choose, switching doors neither gains nor loses you anything; the probability that the car is behind Door A is equal to the probability that it is behind Door B.

But the correct answer, as decades of statistics students have found to their consternation, is that you are twice as likely to win the car if you switch to Door B as you are if you stay with Door A. The reasoning often given for this counterintuitive solution is that, when you originally chose a door, you had a $\frac{1}{3}$ probability of picking the door with the car. Since Monty *always* opens a door with a goat, no matter whether you initially chose the car or not, you have received no new information since then. Therefore, there is still a $\frac{1}{3}$ probability that the door you picked hides the car, and the remaining $\frac{2}{3}$ probability must lie with the only other closed door left.

We can prove this surprising fact using Bayes' rule. Here we have three variables: X, the door chosen by the player; Y, the door behind which the car is hidden; and Z, the door which the host opens. X, Y, and Z can all take the values A, B, or C. We want to prove that $P(Y = B|X = A, Z = C) > P(Y = A|X = A, Z = C)$. Our hypothesis is that the car lies behind Door A; our evidence is that Monty opened Door C. We will leave the proof to the reader—see Study question 1.3.5. To further develop your intuition, you might generalize the game to having 100 doors (which contain 1 hidden car and 99 hidden goats). The contestant still chooses one door, but now Monty opens 98 doors—all revealing goats deliberately—before offering the contestant the chance to switch before the final doors are opened. Now, the choice to switch should be obvious.

Why does Monty opening Door C constitute evidence about the location of the car? It didn't, after all, provide any evidence for whether your initial choice of door was correct. And, surely,

when he was about to open a door, be it *B* or *C*, you knew in advance that you won't find a car behind it. The answer is that there was no way for Monty to open Door *A* after you chose it—but he *could* have opened Door *B*. The fact that he didn't makes it more likely that he opened Door *C* because he was forced to; it provides evidence that the car lies behind Door *B*. This is a general theme of Bayesian analysis: Any hypothesis that has withstood some test of refutation becomes more likely. Door *B* was vulnerable to refutation (i.e., Monty could have opened it), but Door *A* was not. Therefore, Door *B* becomes a more likely location, whereas Door *A* does not.

The reader may find it instructive to note that the explanation above is laden with counter-factual terminology; for example, "He could have opened," "because he was forced," "He was about to open." Indeed, what makes the Monty Hall example unique among probability puzzles is its critical dependence on the process that generated the data. It shows that our beliefs should depend not merely on the facts observed but also on the process that led to those facts. In particular, the information that the car is not behind Door *C*, in itself, is not sufficient to describe the problem; to figure out the probabilities involved, we must also know what options were available to the host before opening Door *C*. In Chapter 4 of this book we will formulate a theory of counterfactuals that will enable us to describe such processes and alternative options, so as to form the correct beliefs about choices.

There is some controversy attached to Bayes' rule. Often, when we are trying to ascertain the probability of a hypothesis given some evidence, we have no way to calculate the prior probability of the hypothesis, $P(A)$, in terms of fractions or frequencies of cases. Consider: If we did not know the proportion of roulette tables to craps tables in the casino, how on Earth could we determine the prior probability $P(\text{"craps"})$? We might be tempted to postulate $P(A) = \frac{1}{2}$ as a way of expressing our ignorance. But what if we have a hunch that roulette tables are less common in this casino, or the tone of the voice of the caller reminds us of a craps dealer we heard yesterday? In cases such as this, in order to use Bayes' rule, we substitute, in place of $P(A)$, our *subjective belief* in the relative truth of the hypothesis compared to other possibilities. The controversy stems from the subjective nature of that belief—how are we to know whether the assigned $P(A)$ accurately summarizes the information we have about the hypothesis? Should we insist on distilling all of our pro and con arguments down to a single number? And even if we do, why should we update our subjective beliefs about hypotheses the same way that we update objective frequencies? Some behavioral experiments suggest that people do not update their beliefs in accordance with Bayes' rule—but many believe that they *should*, and that deviations from the rule represent compromises, if not deficiencies in reasoning, and lead to suboptimal decisions. Debate over the proper use of Bayes' theorem continues to this day. Despite these controversies, however, Bayes' rule is a powerful tool for statistics, and we will use it to great effect throughout this book.

Study questions

Study question 1.3.3

Consider the casino problem described in Section 1.3.6

(a) Compute $P(\text{"craps"}|\text{"11"})$ assuming that there are twice as many roulette tables as craps games at the casino.

(b) *Compute $P("roulette"|"10")$ assuming that there are twice as many craps games as roulette tables at the casino.*

Study question 1.3.4

Suppose we have three cards. Card 1 has two black faces, one on each side; Card 2 has two white faces; and Card 3 has one white face and one black face. You select a card at random and place it on the table. You find that it is black on the face-up side. What is the probability that the face-down side of the card is also black?

(a) *Use your intuition to argue that the probability that the face-down side of the card is also black is $\frac{1}{2}$. Why might it be greater than $\frac{1}{2}$?*

(b) *Express the probabilities and conditional probabilities that you find easy to estimate (for example, $P(C_D = Black)$), in terms of the following variables:*

$$I = \text{Identity of the card selected (Card 1, Card 2, or Card 3)}$$

$$C_D = \text{Color of the face-down side (Black, White)}$$

$$C_U = \text{Color of the face-up side (Black, White)}$$

Find the probability that the face-down side of the selected card is black, using your estimates above.

(c) *Use Bayes' theorem to find the correct probability of a randomly selected card's back being black if you observe that its front is black?*

Study question 1.3.5 (Monty Hall)

Prove, using Bayes' theorem, that switching doors improves your chances of winning the car in the Monty Hall problem.

1.3.8   Expected Values

In statistics, one often deals with data sets and probability distributions that are too large to effectively examine each possible combination of values. Instead, we use statistical measures to represent, with some loss of information, meaningful features of the distribution. One such measure is the *expected value*, also called the *mean*, which can be used when variables take on numerical values. The expected value of a variable X, denoted $E(X)$, is found by multiplying each possible value of the variable by the probability that the variable will take that value, then summing the products:

$$E(X) = \sum_x x\, P(X = x) \tag{1.10}$$

For instance, a variable X representing the outcome of one roll of a fair six-sided die has the following probability distribution: $P(1) = \frac{1}{6}, P(2) = \frac{1}{6}, P(3) = \frac{1}{6}, P(4) = \frac{1}{6}, P(5) = \frac{1}{6}, P(6) = \frac{1}{6}$. The expected value of X is given by:

$$E(X) = \left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{6}\right) + \left(3 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(5 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{6}\right) = 3.5$$

Similarly, the expected value of any function of X—say, $g(X)$—is obtained by summing $g(x)P(X = x)$ over all values of X.

$$E[g(X)] = \sum_x g(x)P(x) \tag{1.11}$$

For example, if after rolling a die, I receive a cash prize equal to the square of the result, we have $g(X) = X^2$, and the expected prize is

$$E[g(X)] = \left(1^2 \times \frac{1}{6}\right) + \left(2^2 \times \frac{1}{6}\right) + \left(3^2 \times \frac{1}{6}\right) + \left(4^2 \times \frac{1}{6}\right) + \left(5^2 \times \frac{1}{6}\right) + \left(6^2 \times \frac{1}{6}\right) = 15.17 \tag{1.12}$$

We can also calculate the expected value of Y conditional on X, $E(Y|X = x)$, by multiplying each possible value y of Y by $P(Y = y|X = x)$, and summing the products.

$$E(Y|X = x) = \sum_y y\, P(Y = y|X = x) \tag{1.13}$$

$E(X)$ is one way to make a "best guess" of X's value. Specifically, out of all the guesses g that we can make, the choice "$g = E(X)$" minimizes the expected square error $E(g - X)^2$. Similarly, $E(Y|X = x)$ represents a best guess of Y, given that we observe $X = x$. If $g = E(Y|X = x)$, then g minimizes the expected square error $E[(g - Y)^2|X = x]$.

For example, the expected age of a 2012 voter, as demonstrated by Table 1.3, is

$$E(Voter's\ Age) = 23.5 \times 0.16 + 37 \times 0.23 + 54.5 \times 0.39 + 70 \times 0.22 = 48.9$$

(For this calculation, we have assumed that every age within each category is equally likely, e.g., a voter is as likely to be 18 as 25, and as likely to be 30 as 44. We have also assumed that the oldest age of any voter is 75.) This means that if we were asked to guess the age of a randomly chosen voter, with the understanding that if we were off by e years, we would lose e^2 dollars, we would lose the least money, on average, if we guessed 48.9. Similarly, if we were asked to guess the age of a random voter younger than the age of 45, our best bet would be

$$E[Voter's\ Age \mid Voter's\ Age < 45] = 23.5 \times 0.40 + 37 \times 0.60 = 31.6 \tag{1.14}$$

The use of expectations as a basis for predictions or "best guesses" hinges to a great extent on an implicit assumption regarding the distribution of X or $Y|X = x$, namely that such distributions are approximately *symmetric*. If, however, the distribution of interest is highly *skewed*, other methods of prediction may be better. In such cases, for example, we might use the median of the distribution of X as our "best guess"; this estimate minimizes the expected absolute error $E(|g - X|)$. We will not pursue such alternative measures further here.

1.3.9 Variance and Covariance

The *variance* of a variable X, denoted $Var(X)$ or σ_X^2, is a measure of roughly how "spread out" the values of X in a data set or population are from their mean. If the values of X all hover close

to one value, the variance will be relatively small; if they cover a large range, the variance will be comparatively large. Mathematically, we define the variance of a variable as the average square difference of that variable from its mean. It can be computed by first finding its mean, μ, and then calculating

$$Var(X) = E((X - \mu)^2) \tag{1.15}$$

The *standard deviation* σ_X of a random variable X is the square root of its variance. Unlike the variance, σ_X is expressed in the same units as X. For example, the variance of under-45 voters' age distribution, according to Table 1.3, can easily be calculated to be (Eq. (1.15)):

$$Var(X) = ((23.5 - 31.5)^2 \times 0.41) + ((37 - 31.5)^2 \times 0.59)$$

$$= (64 \times 0.41) + (30.25 \times .59)$$

$$= 26.24 + 17.85 = 43.09 \text{ years}^2$$

while the standard deviation is

$$\sigma_X = \sqrt{(43.09)} = 6.56 \text{ years}$$

This means that, choosing a voter at random, chances are high that his/her age will fall less than 6.56 years away from the average 31.5. This kind of interpretation can be quantified. For example, for a normally distributed random variable X, approximately two-thirds of the population values of X fall within *one* standard deviation of the expectation, or mean. Further, about 95% fall within *two* standard deviations from the mean.

Of special importance is the expectation of the product $(X - E(X))(Y - E(Y))$, which is known as the *covariance* of X and Y,

$$\sigma_{XY} \triangleq E[(X - E(X))(Y - E(Y))] \tag{1.16}$$

It measures the degree to which X and Y *covary*, that is, the degree to which the two variables vary together, or are "associated." This measure of association actually reflects a specific way in which X and Y covary; it measures the extent to which X and Y *linearly* covary. You can think of this as plotting Y versus X and considering the extent to which a straight *line* captures the way in which Y varies as X changes.

The covariance σ_{XY} is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{1.17}$$

which is a dimensionless number ranging from -1 to 1, which represents the slope of the best-fit line after we normalize both X and Y by their respective standard deviations. ρ_{XY} is one if and only if one variable can predict the other in a *linear* fashion, and it is zero whenever such a linear prediction is no better than a random guess. The significance of σ_{XY} and ρ_{XY} will be discussed in the next section. At this point, it is sufficient to note that these degrees of covariation can be readily computed from the joint distribution $P(x, y)$, using Eqs. (1.16) and (1.17). Moreover, both σ_{XY} and ρ_{XY} vanish when X and Y are independent. Note that nonlinear relationships between Y and X cannot naturally be captured by a simple numerical summary; they require a full specification of the conditional probability $P(Y = y|X = x)$.

This result is not surprising, since Y (the sum of the two dice) can be written as

$$Y = X + Z$$

where Z is the outcome of Die 2, and it stands to reason that if X increases by one unit, say from $X = 3$ to $X = 4$, then $E[Y]$ will, likewise, increase by one unit. The reader might be a bit surprised, however, to find out that the reverse is not the case; the regression of X on Y does not have a slope of 1.0. To see why, we write

$$E[X|Y = y] = E[Y - Z|Y = y] = 1.0y - E[Z|Y = y] \tag{1.20}$$

and realize that the added term, $E[Z|Y = y]$, since it depends (linearly) on y, makes the slope less than unity. We can in fact compute the exact value of $E[X|Y = y]$ by appealing to symmetry and write

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

which gives, after substituting in Eq. (1.20),

$$E[X|Y = y] = 0.5y$$

The reason for this reduction is that, when we increase Y by one unit, each of X and Z contributes equally to this increase on average. This matches intuition; observing that the sum of the two dice is $Y = 10$, our best estimate of each is $X = 5$ and $Z = 5$.

In general, if we write the regression equation for Y on X as

$$y = a + bx \tag{1.21}$$

the slope b is denoted by R_{YX}, and it can be written in terms of the covariate σ_{XY} as follows:

$$b = R_{YX} = \frac{\sigma_{XY}}{\sigma_X^2} \tag{1.22}$$

From this equation, we see clearly that the slope of Y on X may differ from the slope of X on Y—that is, in most cases, $R_{YX} \neq R_{XY}$. ($R_{YX} = R_{XY}$ only when the variance of X is equal to the variance of Y.) The slope of the regression line can be positive, negative, or zero. If it is positive, X and Y are said to have a *positive correlation*, meaning that as the value of X gets higher, the value of Y gets higher; if it is negative, X and Y are said to have a *negative correlation*, meaning that as the value of X gets higher, the value of Y gets lower; if it is zero (a horizontal line), X and Y have no linear correlation, and knowing the value of X does not assist us in predicting the value of Y, at least linearly. If two variables are correlated, whether positively or negatively (or in some other way), they are dependent.

1.3.11 Multiple Regression

It is also possible to regress a variable on several variables, using *multiple linear regression*. For instance, if we wanted to predict the value of a variable Y using the values of the variables X and Z, we could perform multiple linear regression of Y on $\{X, Z\}$, and estimate a regression relationship

$$y = r_0 + r_1 x + r_2 z \tag{1.23}$$

which represents an inclined plane through the three-dimensional coordinate system.

We can create a three-dimensional scatter plot, with values of Y on the y-axis, X on the x-axis, and Z on the z-axis. Then, we can cut the scatter plot into slices along the Z-axis. Each slice will constitute a two-dimensional scatter plot of the kind shown in Figure 1.4. Each of those 2-D scatter plots will have a regression line with a slope r_1. Slicing along the X-axis will give the slope r_2.

The slope of Y on X when we hold Z constant is called the *partial regression coefficient* and is denoted by $R_{YX \cdot Z}$. Note that it is possible for R_{YX} to be positive, whereas $R_{YX \cdot Z}$ is negative as shown in Figure 1.1. This is a manifestation of Simpson's Paradox: positive association between Y and X overall, that becomes negative when we condition on the third variable Z.

The computation of partial regression coefficients (e.g., r_1 and r_2 in (1.23)) is greatly facilitated by a theorem that is one of the most fundamental results in regression analysis. It states that if we write Y as a linear combination of variables $X_1, X_2, \ldots, X_k$ plus a noise term ϵ,

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \qquad (1.24)$$

then, regardless of the underlying distribution of $Y, X_1, X_2, \ldots, X_k$, the best least-square coefficients are obtained when ϵ is uncorrelated with each of the regressors $X_1, X_2, \ldots, X_k$. That is,

$$Cov(\epsilon, X_i) = 0 \quad \text{for} \quad i = 1, 2, \ldots, k$$

To see how this *orthogonality principle* is used to our advantage, assume we wish to compute the best estimate of $X = Die\ 1$ given the sum

$$Y = Die\ 1 + Die\ 2$$

Writing

$$X = \alpha + \beta Y + \epsilon \qquad (1.25a)$$

our goal is to find α and β in terms of estimable statistical measures. Assuming without loss of generality $E[\epsilon] = 0$, and taking expectation on both sides of the equation, we obtain

$$E[X] = \alpha + \beta E[Y] \qquad (1.25b)$$

Further multiplying both sides of the equation by Y and taking the expectation gives

$$E[XY] = \alpha E[Y] + \beta E[Y^2] + E[Y\epsilon] \qquad (1.26)$$

The orthogonality principle dictates $E[Y\epsilon] = 0$, and (1.25) and (1.26) yield two equations with two unknowns, α and β. Solving for α and β, we obtain

$$\alpha = E(X) - E(Y)\frac{\sigma_{XY}}{\sigma_Y^2}$$

$$\beta = \frac{\sigma_{XY}}{\sigma_Y^2}$$

which completes the derivation. The slope β could have been obtained from Eq. (1.22), by simply reversing X and Y, but the derivation above demonstrates a general method of computing slopes, in two or more dimensions.

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

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

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

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 1.3.8. [Hint: Apply Eq. (1.27) and use the variances and covariances computed for part (a) of 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

object. A mathematical graph is a collection of *vertices* (or, as we will call them, *nodes*) and edges. The nodes in a graph are connected (or not) by the edges. Figure 1.5 illustrates a simple graph. X, Y, and Z (the dots) are nodes, and A and B (the lines) are edges.
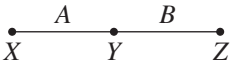
$$\underset{X}{\bullet} \overset{A}{\rule{2cm}{0.4pt}} \underset{Y}{\bullet} \overset{B}{\rule{2cm}{0.4pt}} \underset{Z}{\bullet}$$

Figure 1.5   An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

Two nodes are *adjacent* if there is an edge between them. In Figure 1.5, X and Y are adjacent, and Y and Z are adjacent. A graph is said to be a *complete graph* if there is an edge between every pair of nodes in the graph.

A *path* between two nodes X and Y is a sequence of nodes beginning with X and ending with Y, in which each node is connected to the next by an edge. For instance, in Figure 1.5, there is a path from X to Z, because X is connected to Y, and Y is connected to Z.

Edges in a graph can be *directed* or *undirected*. Both of the edges in Figure 1.5 are undirected, because they have no designated "in" and "out" ends. A directed edge, on the other hand, goes out of one node and into another, with the direction indicated by an arrow head. A graph in which all of the edges are directed is a *directed graph*. Figure 1.6 illustrates a directed graph. In Figure 1.6, A is a directed edge from X to Y and B is a directed edge from Y to Z.

$$\underset{X}{\bullet} \overset{A}{\longrightarrow} \underset{Y}{\bullet} \overset{B}{\longrightarrow} \underset{Z}{\bullet}$$

Figure 1.6   A directed graph in which node X is a parent of Y and Y is a parent of Z

The node that a directed edge starts from is called the *parent* of the node that the edge goes into; conversely, the node that the edge goes into is the *child* of the node it comes from. In Figure 1.6, X is the parent of Y, and Y is the parent of Z; accordingly, Y is the child of X, and Z is the child of Y. A path between two nodes is a *directed* path if it can be traced along the arrows, that is, if no node on the path has two edges on the path directed into it, or two edges directed out of it. If two nodes are connected by a directed path, then the first node is the *ancestor* of every node on the path, and every node on the path is the *descendant* of the first node. (Think of this as an analogy to parent nodes and child nodes: parents are the ancestors of their children, and of their children's children, and of their children's children's children, etc.) For instance, in Figure 1.6, X is the ancestor of both Y and Z, and both Y and Z are descendants of X.

When a directed path exists from a node to itself, the path (and graph) is called *cyclic*. A directed graph with no cycles is *acyclic*. For example, in Figure 1.7(a) the graph is acyclic; however, the graph in Figure 1.7(b) is cyclic. Note that in 1.7(a) there is no directed path from any node to itself, whereas in 1.7(b) there are directed paths from X back to X, for example.

Because there are edges connecting Z to X and Y, we can conclude just by looking at the graphical model that there is some function f_Z in the model that assigns Z a value based on X and Y, and therefore that X and Y are causes of Z. However, without the fuller specification of an SCM, we can't tell from the graph what the function is that defines Z—or, in other words, *how X and Y cause Z*.

If graphical models contain less information than SCMs, why do we use them at all? There are several reasons. First, usually the knowledge that we have about causal relationships is not quantitative, as demanded by an SCM, but qualitative, as represented in a graphical model. We know off-hand that sex is a cause of height and that height is a cause of performance in basketball, but we would hesitate to give numerical values to these relationships. We could, instead of drawing a graph, simply create a partially specified version of the SCM:

SCM 1.5.2   (Basketball Performance Based on Height and Sex)

$$V = \{\text{Height, Sex, Performance}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f1, f2\}$$

$$\text{Sex} = U_1$$

$$\text{Height} = f_1(\text{Sex}, U_2)$$

$$\text{Performance} = f_2(\text{Height, Sex}, U_3)$$

Here, $U = \{U_1, U_2, U_3\}$ represents unmeasured factors that we do not care to name, but that affect the variables in V that we can measure. The U factors are sometimes called "error terms" or "omitted factors." These represent additional unknown and/or random exogenous causes of what we observe.

But graphical models provide a more intuitive understanding of causality than do such partially specified SCMs. Consider the SCM and its associated graphical model introduced above; while the SCM and its graphical model contain the same information, that is, that X causes Z and Y causes Z, that information is more quickly and easily ascertained by looking at the graphical model.

Study questions

Study question 1.5.1

Suppose we have the following SCM. Assume all exogenous variables are independent and that the expected value of each is 0.

SCM 1.5.3

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

$$f_X : X = U_X$$

$$f_Y : Y = \frac{X}{3} + U_Y$$

$$f_Z : Z = \frac{Y}{16} + U_Z$$

(a) *Draw the graph that complies with the model.*
(b) *Determine the best guess of the value (expected value) of Z, given that we observe Y = 3.*
(c) *Determine the best guess of the value of Z, given that we observe X = 3.*
(d) *Determine the best guess of the value of Z, given that we observe X = 1 and Y = 3.*
(e) *Assume that all exogenous variables are normally distributed with zero means and unit variance, that is, σ = 1.*
    (i) *Determine the best guess of X, given that we observed Y = 2.*
    (ii) *(Advanced) Determine the best guess of Y, given that we observed X = 1 and Z = 3. [Hint: You may wish to use the technique of multiple regression, together with the fact that, for every three normally distributed variables, say X, Y, and Z, we have $E[Y|X = x, Z = z] = R_{YX \cdot Z}x + R_{YZ \cdot X}z.$]*
(f) *Determine the best guess of the value of Z, given that we know X = 3.*

1.5.2 Product Decomposition

Another advantage of graphical models is that they allow us to express joint distributions very efficiently. So far, we have presented joint distributions in two ways. First, we have used tables, in which we assigned a probability to every possible combination of values. This is intuitively easy to parse, but in models with many variables, it can take up a prohibitive amount of space; 10 binary variables would require a table with 1024 rows!

Second, in a fully specified SCM, we can represent the joint distributions of *n* variables with greater efficiency: We need only to specify the *n* functions that govern the relationships between the variables, and then from the probabilities of the error terms, we can discover all the probabilities that govern the joint distribution. But we are not always in a position to fully specify a model; we may know that one variable is a cause of another but not the form of the equation relating them, or we may not know the distributions of the error terms. Even if we know these objects, writing them down may be easier said than done, especially when the variables are discrete and the functions do not have familiar algebraic expressions.

Fortunately, we can use graphical models to help overcome both of these barriers through the following rule.

Rule of product decomposition

For any model whose graph is acyclic, the joint distribution of the variables in the model is given by the product of the conditional distributions $P(child|parents)$ over all the "families" in the graph. Formally, we write this rule as

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i|pa_i) \tag{1.29}$$

where pa_i stands for the values of the parents of variable X_i, and the product $\prod_i$ runs over all *i*, from 1 to *n*. The relationship (1.29) follows from certain universally true independencies among the variables, which will be discussed in the next chapter in more detail.

For example, in a simple chain graph $X \rightarrow Y \rightarrow Z$, we can write directly:

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

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 $q_2 = P(x_1|z_1)$.

(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 $P(y_1|x_1) - P(y_1|x_o)$ for three populations: (1) those carrying the syndrome, (2) those not carrying the syndrome, and (3) the population as a whole.*
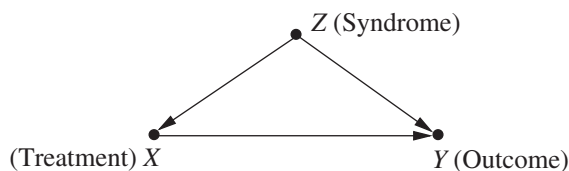

Figure 1.10   Model showing an unobserved syndrome, Z, affecting both treatment (X) and outcome (Y)

(c) Using your results for (b), find a combination of parameters that exhibits Simpson's reversal.

Study question 1.5.3

Consider a graph $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ of binary random variables, and assume that the conditional probabilities between any two consecutive variables are given by

$$P(X_i = 1 | X_{i-1} = 1) = p$$

$$P(X_i = 1 | X_{i-1} = 0) = q$$

$$P(X_1 = 1) = p_0$$

Compute the following probabilities

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0)$$

$$P(X_4 = 1 | X_1 = 1)$$

$$P(X_1 = 1 | X_4 = 1)$$

$$P(X_3 = 1 | X_1 = 0, X_4 = 1)$$

Study question 1.5.4

Define the structural model that corresponds to the Monty Hall problem, and use it to describe the joint distribution of all variables.

Bibliographical Notes for Chapter 1

An extensive account of the history of Simpson's paradox is given in Pearl (2009, pp. 174–182), including many attempts by statisticians to resolve it without invoking causation. A more recent account, geared for statistics instructors is given in (Pearl 2014b). Among the many texts that provide basic introductions to probability theory, Lindley (2014) and Pearl (1988, Chapters 1 and 2) are the closest in spirit to the Bayesian perspective used in Chapter 1. The textbooks by Selvin (2004) and Moore et al. (2014) provide excellent introductions to classical methods of statistics, including parameter estimation, hypothesis testing and regression analysis.

The Monty Hall problem, discussed in Section 1.3, appears in many introductory books on probability theory (e.g., Grinstead and Snell 1998, p. 136; Lindley 2014, p. 201) and is mathematically equivalent to the "Three Prisoners Dilemma" discussed in (Pearl 1988, pp. 58–62). Friendly introductions to graphical models are given in Elwert (2013), Glymour and Greenland (2008), and the more advanced texts of Pearl (1988, Chapter 3), Lauritzen (1996) and Koller and Friedman (2009). The product decomposition rule of Section 1.5.2 was used in Howard and Matheson (1981) and Kiiveri et al. (1984) and became the semantic

SCM 2.2.3 (Work Hours, Training, and Race Time)

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

$$f_X : X = U_X$$

$$f_Y : Y = 84 - x + U_Y$$

$$f_Z : Z = \frac{100}{y} + U_Z$$

SCMs 2.2.1–2.2.3 share the graphical model shown in Figure 2.1.

SCMs 2.2.1 and 2.2.3 deal with continuous variables; SCM 2.2.2 deals with categorical variables. The relationships between the variables in 2.2.1 are all positive (i.e., the higher the value of the parent variable, the higher the value of the child variable); the correlations between the variables in 2.2.3 are all negative (i.e., the higher the value of the parent variable, the lower the value of the child variable); the correlations between the variables in 2.2.2 are not linear at all, but logical. No two of the SCMs share any functions in common. But because they share a common graphical structure, the data sets generated by all three SCMs must share certain independencies—and we can predict those independencies simply by examining the graphical model in Figure 2.1. The independencies shared by data sets generated by these three SCMs, and the dependencies that are likely shared by all such SCMs, are these:

1. ***Z and Y are likely dependent***
   For some $z, y, P(Z = z | Y = y) \neq P(Z = z)$
2. ***Y and X are likely dependent***
   For some $y, x, P(Y = y | X = x) \neq P(Y = y)$
3. ***Z and X are likely dependent***
   For some $z, x, P(Z = z | X = x) \neq P(Z = z)$
4. ***Z and X are independent, conditional on Y***
   For all $x, y, z, P(Z = z | X = x, Y = y) = P(Z = z | Y = y)$

To understand why these independencies and dependencies hold, let's examine the graphical model. First, we will verify that any two variables with an edge between them are likely dependent. Remember that an arrow from one variable to another indicates that the first variable causes the second—that is, the value of the first variable is part of the function that determines the value of the second. Therefore, the second variable *depends* on the first for
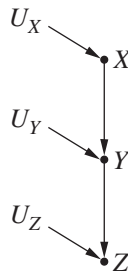

Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

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 ($P(X=1)=P(U_X)=1/2$); they are rare, and can be ignored for all practical purposes.

Rule 1 (Conditional Independence in Chains) *Two variables, X and Y, are conditionally independent given Z, if there is only one unidirectional path between X and Y and Z is any set of variables that intercepts that path.*

An important note: Rule 1 only holds when we assume that the error terms U_X, U_Y, and U_Z are independent of each other. If, for instance, U_X were a cause of U_Y, then conditioning on Z would not necessarily make X and Y independent—because variations in X could still be associated with variations in Y, through their error terms.

Now, consider the graphical model in Figure 2.2. This structure might represent, for example, the causal mechanism that connects a day's temperature in a city in degrees Fahrenheit (X), the number of sales at a local ice cream shop on that day (Y), and the number of violent crimes in the city on that day (Z). Possible functional relationships between these variables are given in SCM 2.2.5. Or the structure might represent, as in SCM 2.2.6, the causal mechanism that connects the state (up or down) of a switch (X), the state (on or off) of one light bulb (Y), and the state (on or off) of a second light bulb (Z). The exogenous variables U_X, U_Y, and U_Z represent other, possibly random, factors that influence the operation of these devices.

SCM 2.2.5 (Temperature, Ice Cream Sales, and Crime)

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

$$f_X : X = U_X$$

$$f_Y : Y = 4x + U_Y$$

$$f_Z : Z = \frac{x}{10} + U_Z$$

SCM 2.2.6 (Switch and Two Light Bulbs)

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

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} \text{On IF } (X = \text{Up AND } U_Y = 0) \text{ OR } (X = \text{Down AND } U_Y = 1) \\ \text{Off otherwise} \end{cases}$$

$$f_Z : Z = \begin{cases} \text{On IF } (X = \text{Up AND } U_Z = 0) \text{ OR } (X = \text{Down AND } U_Z = 1) \\ \text{Off otherwise} \end{cases}$$
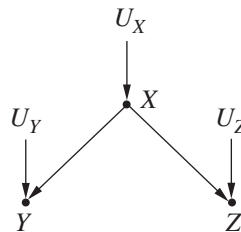

Figure 2.2 The graphical model of SCMs 2.2.5 and 2.2.6

If we assume that the error terms U_X, U_Y, and U_Z are independent, then by examining the graphical model in Figure 2.2, we can determine that SCMs 2.2.5 and 2.2.6 share the following dependencies and independencies:

1. **X and Y are likely dependent.**
   For some x, y, $P(X = x|Y = y) \neq P(X = x)$
2. **X and Z are likely dependent.**
   For some x, z, $P(X = x|Z = z) \neq P(X = x)$
3. **Z and Y are likely dependent.**
   For some z, y, $P(Z = z|Y = y) \neq P(Z = z)$
4. **Y and Z are independent, conditional on X.**
   For all x, y, z, $P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

Points 1 and 2 follow, once again, from the fact that Y and Z are both directly connected to X by an arrow, so when the value of X changes, the values of both Y and Z likely change. This tells us something further, however: If Y changes when X changes, and Z changes when X changes, then it is likely (though not certain) that Y changes together with Z, and vice versa. Therefore, since a change in the value of Y gives us information about an associated change in the value of Z, Y and Z are likely dependent variables.

Why, then, are Y and Z independent conditional on X? Well, what happens when we condition on X? We filter the data based on the value of X. So now, we're only comparing cases where the value of X is constant. Since X does not change, the values of Y and Z do not change in accordance with it—they change only in response to U_Y and U_Z, which we have assumed to be independent. Therefore, any additional changes in the values of Y and Z must be independent of each other.

This configuration of variables—three nodes, with two arrows emanating from the middle variable—is called a *fork*. The middle variable in a fork is the *common cause* of the other two variables, and of any of their descendants. If two variables share a common cause, and if that common cause is part of the only path between them, then analogous reasoning to the above tells us that these dependencies and conditional independencies are true of those variables. Therefore, we come by another rule:

Rule 2 (Conditional Independence in Forks) *If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.*

2.3 Colliders

So far we have looked at two simple configurations of edges and nodes that can occur on a path between two variables: chains and forks. There is a third such configuration that we speak of separately, because it carries with it unique considerations and challenges. The third configuration contains a *collider* node, and it occurs when one node receives edges from two other nodes. The simplest graphical causal model containing a collider is illustrated in Figure 2.3, representing a common effect, Z, of two causes X and Y.

As is the case with every graphical causal model, all SCMs that have Figure 2.3 as their graph share a set of dependencies and independencies that we can determine from the graphical
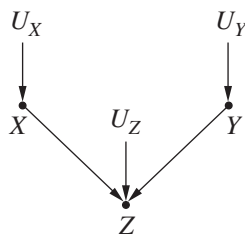
Figure 2.3   A simple collider

model alone. In the case of the model in Figure 2.3, assuming independence of U_X, U_Y, and U_Z, these independencies are as follows:

1. **X and Z are** *likely* **dependent.**
   For some $x, z, P(X = x | Z = z) \neq P(X = x)$
2. **Y and Z are** *likely* **dependent.**
   For some $y, z, P(Y = y | Z = z) \neq P(Y = y)$
3. **X and Y are independent.**
   For all $x, y, P(X = x | Y = y) = P(X = x)$
4. **X and Y are** *likely* **dependent conditional on Z.**
   For some $x, y, z, P(X = x | Y = y, Z = z) \neq P(X = x | Z = z)$

The truth of the first two points was established in Section 2.2. Point 3 is self-evident; neither X nor Y is a descendant or an ancestor of the other, nor do they depend for their value on the same variable. They respond only to U_X and U_Y, which are assumed independent, so there is no causal mechanism by which variations in the value of X should be associated with variations in the value of Y. This independence also reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.

Why, then, does point 4 hold? Why would two independent variables suddenly become dependent when we condition on their common effect? To answer this question, we return again to the definition of conditioning as filtering by the value of the conditioning variable. When we condition on Z, we limit our comparisons to cases in which Z takes the same value. But remember that Z depends, for its value, on X and Y. So, when comparing cases where Z ~~tak~~ takes some ~~lle, the~~ value, any change in value of X must be compensated for by a change in the value of Y—otherwise, the value of Z would change as well.

The reasoning behind this attribute of colliders—that conditioning on a collision node produces a dependence between the node's parents—can be difficult to grasp at first. In the most basic situation where $Z = X + Y$, and X and Y are independent variables, we have the following logic: If I tell you that $X = 3$, you learn nothing about the potential value of Y, because the two numbers are independent. On the other hand, if I start by telling you that $Z = 10$, then telling you that $X = 3$ immediately tells you that Y must be 7. Thus, X and Y are dependent, *given that $Z = 10$.*

This phenomenon can be further clarified through a real-life example. For instance, suppose a certain college gives scholarships to two types of students: those with unusual musical talents and those with extraordinary grade point averages. Ordinarily, musical talent and scholastic achievement are independent traits, so, in the population at large, finding a person with musical

talent tells us nothing about that person's grades. However, discovering that a person is on a scholarship changes things; knowing that the person lacks musical talent then tells us immediately that he is likely to have high grade point average. Thus, two variables that are marginally independent become dependent upon learning the value of a third variable (scholarship) that is a common effect of the first two.

Let's examine a numerical example. Consider a simultaneous (independent) toss of two fair coins and a bell that rings whenever at least one of the coins lands on heads. Let the outcomes of the two coins be denoted X and Y, respectively, and let Z stand for the state of the bell, with $Z = 1$ representing ringing, and $Z = 0$ representing silence. This mechanism can be represented as a collider as in Figure 2.3, in which the outcomes of the two coins are the parent nodes, and the state of the bell is the collision node.

If we know that Coin 1 landed on heads, it tells us nothing about the outcome of Coin 2, due to their independence. But suppose that we hear the bell ring and then we learn that Coin 1 landed on tails. We now know that Coin 2 must have landed on heads. Similarly, if we assume that we've heard the bell ring, the probability that Coin 1 landed on heads changes if we learn that Coin 2 also landed on heads. This particular change in probability is somewhat subtler than the first case.

To see the latter calculation, consider the initial probabilities as shown in Table 2.1.

We see that

$$P(X = \text{"Heads"}|Y = \text{"Heads"}) = P(X = \text{"Tails"}|Y = \text{"Tails"}) = \frac{1}{2}$$

That is, X and Y are independent. Now, let's condition on $Z = 1$ and $Z = 0$ (the bell ringing and not ringing). The resulting data subsets are shown in Table 2.2.

By calculating the probabilities in these tables, we obtain

$$P(X = \text{"Heads"}|Z = 1) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$$

If we further filter the $Z = 1$ subtable to examine only those cases where $Y = \text{"Heads"}$, we get

$$P(X = \text{"Heads"}|Y = \text{"Heads"}, Z = 1) = \frac{1}{2}$$

We see that, given $Z = 1$, the probability of $X = \text{"Heads"}$ changes from $\frac{2}{3}$ to $\frac{1}{2}$ upon learning that $Y = \text{"Heads."}$ So, clearly, X and Y are dependent given $Z = 1$. A more pronounced dependence occurs, of course, when the bell does not ring ($Z = 0$), because then we know that both coins must have landed on tails.

Table 2.1 Probability distribution for two flips of a fair coin, with X representing flip one, Y representing flip two, and Z representing a bell that rings if either flip results in heads

X	Y	Z	$P(X, Y, Z)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	0	0.25

Table 2.2 Conditional probability distributions for the distribution in Table 2.1. (Top: Distribution conditional on $Z = 1$. Bottom: Distribution conditional on $Z = 0$)

X	Y	$P(X, Y | Z = 1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0

X	Y	$Pr(X, Y | Z = 0)$
Heads	Heads	0
Heads	Tails	0
Tails	Heads	0
Tails	Tails	1

Another example of colliders in action—one that may serve to further illuminate the difficulty that such configurations can present to statisticians—is the Monty Hall Problem, which we first encountered in Section 1.3. At its heart, the Monty Hall Problem reflects the presence of a collider. Your initial choice of door is one parent node; the door behind which the car is placed is the other parent node; and the door Monty opens to reveal a goat is the collision node, causally affected by both the other two variables. The causation here is clear: If you choose Door *A*, and if Door *A* has a goat behind it, Monty is forced to open whichever of the remaining doors that has a goat behind it.

Your initial choice and the location of the car are independent; that's why you initially have a $\frac{1}{3}$ chance of choosing the door with the car behind it. However, as with the two independent coins, conditional on Monty's choice of door, your initial choice and the placement of the prizes are dependent. Though the car may only be behind Door *B* in $\frac{1}{3}$ of cases, it will be behind Door *B* in $\frac{2}{3}$ of cases in which you choose Door *A* and Monty opened Door *C*.

Just as conditioning on a collider makes previously independent variables dependent, so too does conditioning on any descendant of a collider. To see why this is true, let's return to our example of two independent coins and a bell. Suppose we do not hear the bell directly, but instead rely on a witness who is somewhat unreliable; whenever the bell *does not ring*, there is 50% chance that our witness will falsely report that it did. Letting *W* stand for the witness's report, the causal structure is shown in Figure 2.4, and the probabilities for all combinations of *X*, *Y*, and *W* are shown in Table 2.3.

The reader can easily verify that, based on this table, we have

$$P(X = \text{``Heads''} | Y = \text{``Heads''}) = P(X = \text{``Heads''}) = \frac{1}{2}$$

and

$$P(X = \text{``Heads''} | W = 1) = (0.25 + 0.25) \div (0.25 + 0.25 + 0.25 + 0.125) = \frac{0.5}{0.875}$$

and

$$P(X = \text{``Heads''} | Y = \text{``Heads''}, W = 1) = 0.25 \div (0.25 + 0.25) = 0.5 < \frac{0.5}{0.875}$$
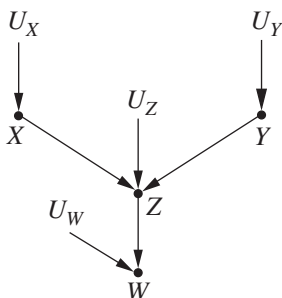
Figure 2.4   A simple collider, Z, with one child, W, representing the scenario from Table 2.3, with X representing one coin flip, Y representing the second coin flip, Z representing a bell that rings if either X or Y is heads, and W representing an unreliable witness who reports on whether or not the bell has rung

Table 2.3   Probability distribution for two flips of a fair coin and a bell that rings if either flip results in heads, with X representing flip one, Y representing flip two, and W representing a witness who, with variable reliability, reports whether or not the bell has rung

X	Y	W	$P(X, Y, W)$
Heads	Heads	1	0.25
Heads	Tails	1	0.25
Tails	Heads	1	0.25
Tails	Tails	1	0.125
Tails	Tails	0	0.125

Thus, X and Y are independent before reading the witness report, but become dependent thereafter.

These considerations lead us to a third rule, in addition to the two we established in Section 2.2.

Rule 3 (Conditional Independence in Colliders) *If a variable Z is the collision node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.*

Rule 3 is extremely important to the study of causality. In the coming chapters, we will see that it allows us to test whether a causal model could have generated a data set, to discover models from data, and to fully resolve Simpson's Paradox by determining which variables to measure and how to estimate causal effects under confounding.

Remark   Inquisitive students may wonder why it is that dependencies associated with conditioning on a collider are so surprising to most people—as in, for example, the Monty Hall example. The reason is that humans tend to associate dependence with causation. Accordingly, they assume (wrongly) that statistical dependence between two variables can only exist if there is a causal mechanism that generates such dependence; that is, either one of the variables causes the other or a third variable causes both. In the case of a collider, they are surprised to find a

$$X \longrightarrow R \longrightarrow S \longrightarrow T \longleftarrow U \longleftarrow V \longrightarrow Y$$

Figure 2.5 A directed graph for demonstrating conditional independence (error terms are not shown explicitly)

$$X \longrightarrow R \longrightarrow S \longrightarrow T \longleftarrow U \longleftarrow V \longrightarrow Y$$
$$\downarrow$$
$$P$$

Figure 2.6 A directed graph in which P is a descendant of a collider

dependence that is created in a third way, thus violating the assumption of "no correlation without causation."

Study questions

Study question 2.3.1

(a) *List all pairs of variables in Figure 2.5 that are independent conditional on the set $Z = \{R, V\}$.*

(b) *For each pair of nonadjacent variables in Figure 2.5, give a set of variables that, when conditioned on, renders that pair independent.*

(c) *List all pairs of variables in Figure 2.6 that are independent conditional on the set $Z = \{R, P\}$.*

(d) *For each pair of nonadjacent variables in Figure 2.6, give a set of variables that, when conditioned on, renders that pair independent.*

(e) *Suppose we generate data by the model described in Figure 2.5, and we fit them with the linear equation $Y = a + bX + cZ$. Which of the variables in the model may be chosen for Z so as to guarantee that the slope b would be equal to zero? [Hint: Recall, a non zero slope implies that Y and X are dependent given Z.]*

(f) *Continuing question (e), suppose we fit the data with the equation:* but now in reference to Figure 2.6,

$$Y = a + bX + cR + dS + eT + fP$$

which of the coefficients would be zero?

2.4 *d*-separation

Causal models are generally not as simple as the cases we have examined so far. Specifically, it is rare for a graphical model to consist of a single path between variables. In most graphical models, pairs of variables will have multiple possible paths connecting them, and each such a path will traverse a variety of chains, forks, and colliders. The question remains whether there is a criterion or process that can be applied to a graphical causal model of *any* complexity in order to predict dependencies that are shared by all data sets generated by that graph.

There is, indeed, such a process: *d-separation*, which is built upon the rules established in the previous section. *d*-separation (the *d* stands for "directional") allows us to determine, for any pair of nodes, whether the nodes are *d-connected*, meaning there exists a connecting path between them, or *d-separated*, meaning there exists no such path. When we say that a pair of nodes are *d*-separated, we mean that the variables they represent are definitely independent; when we say that a pair of nodes are *d*-connected, we mean that they are possibly, or most likely, dependent.[1]

Two nodes *X* and *Y* are *d*-separated if every path between them (should any exist) is *blocked*. If even one path between *X* and *Y* is unblocked, *X* and *Y* are *d*-connected. The paths between variables can be thought of as pipes, and dependence as the water that flows through them; if even one pipe is unblocked, some water can pass from one place to another, and if a single path is clear, the variables at either end will be dependent. However, a pipe need only be blocked in one place to stop the flow of water through it, and similarly, it takes only one node to block the passage of dependence in an entire path.

There are certain kinds of nodes that can block a path, depending on whether we are performing unconditional or conditional *d*-separation. If we are not conditioning on any variable, then only colliders can block a path. The reasoning for this is fairly straightforward: as we saw in Section 2.3, unconditional dependence can't pass through a collider. So if every path between two nodes *X* and *Y* has a collider in it, then *X* and *Y* cannot be unconditionally dependent; they must be marginally independent.

If, however, we are conditioning on a set of nodes *Z*, then the following kinds of nodes can block a path:

- A collider that is not conditioned on (i.e., not in *Z*), and that has no descendants in *Z*.
- A chain or fork whose middle node is in *Z*.

The reasoning behind these points goes back to what we learned in Sections 2.2 and 2.3. A collider does not allow dependence to flow between its parents, thus blocking the path. But Rule 3 tells us that when we condition on a collider or its descendants, the parent nodes may become dependent. So a collider whose collision node is not in the conditioning set *Z* would block dependence from passing through a path, but one whose collision node, or its descendants, *is* in the conditioning set would not. Conversely, dependence can pass through noncolliders—chains and forks—but Rules 1 and 2 tell us that when we condition on them, the variables on either end of those paths become independent (when we consider one path at a time). So any noncollision node in the conditioning set would block dependence, whereas one that is not in the conditioning set would allow dependence through.

We are now prepared to give a general definition of *d*-separation:

Definition 2.4.1 (***d*-separation**) *A path p is blocked by a set of nodes Z if and only if*

1. *p contains a chain of nodes A → B → C or a fork A ← B → C such that the middle node B is in Z (i.e., B is conditioned on), or*
2. *p contains a collider A → B ← C such that the collision node B is not in Z, and no descendant of B is in Z.*

---

[1] The *d*-connected variables will be dependent for almost all sets of functions assigned to arrows in the graph, the exception being the sorts of intransitive cases discussed in Section 2.2.
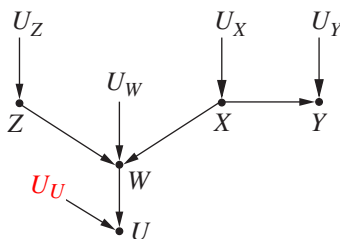
Figure 2.7   A graphical model containing a collider with child and a fork

If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z.

Armed with the tool of *d*-separation, we can now look at some more complex graphical models and determine which variables in them are independent and dependent, both marginally and conditional on other variables. Let's take, for example, the graphical model in Figure 2.7. This graph might be associated with any number of causal models. The variables might be discrete, continuous, or a mixture of the two; the relationships between them might be linear, exponential, or any of an infinite number of other relations. No matter the model, however, *d*-separation will always provide the same set of independencies in the data the model generates.

In particular, let's look at the relationship between *Z* and *Y*. Using an empty conditioning set, they are *d*-separated, which tells us that *Z* and *Y* are unconditionally independent. Why? Because there is no unblocked path between them. There is only one path between *Z* and *Y*, and that path is blocked by a collider ($Z \rightarrow W \leftarrow X$).

But suppose we condition on *W*. *d*-separation tells us that *Z* and *Y* are *d*-connected, conditional on *W*. The reason is that our conditioning set is now {*W*}, and since the only path between *Z* and *Y* contains a fork (*X*) that is not in that set, and the only collider (*W*) on the path is in that set, that path is not blocked. (Remember that conditioning on colliders "unblocks" them.) The same is true if we condition on *U*, because *U* is a descendant of a collider along the path between *Z* and *Y*.

On the other hand, if we condition on the set {*W*, *X*}, *Z* and *Y* remain independent. This time, the path between *Z* and *Y* is blocked by the first criterion, rather than the second: There is now a noncollider node (*X*) on the path that is in the conditioning set. Though *W* has been unblocked by conditioning, one blocked node is sufficient to block the entire path. Since the only path between *Z* and *Y* is blocked by this conditioning set, *Z* and *Y* are *d*-separated conditional on {*W*, *X*}.

Now, consider what happens when we add another path between *Z* and *Y*, as in Figure 2.8. *Z* and *Y* are now unconditionally dependent. Why? Because there is a path between them ($Z \leftarrow T \rightarrow Y$) that contains no colliders. If we condition on *T*, however, that path is blocked, and *Z* and *Y* become independent again. Conditioning on {*T*, *W*}, on the other hand, makes them *d*-connected again (conditioning on *T* blocks the path $Z \leftarrow T \rightarrow Y$, but conditioning on *W* unblocks the path $Z \rightarrow W \leftarrow X \rightarrow Y$). And if we add *X* to the conditioning set, making it {*T*, *W*, *X*}, *Z*, and *Y* become independent yet again! In this graph, *Z* and *Y* are *d*-connected (and therefore likely dependent) conditional
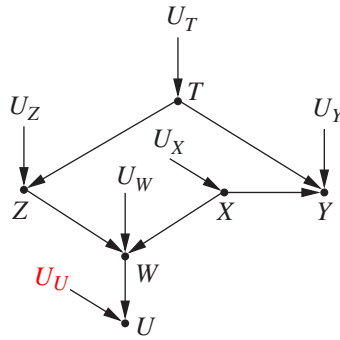
Figure 2.8   The model from Figure 2.7 with an additional forked path between *Z* and *Y*

on $W, U, \{W, U\}, \{W, T\}, \{U, T\}, \{W, U, T\}, \{W, X\}, \{U, X\}$,   and $\{W, U, X\}$. They are
d-separated (and therefore independent) conditional on $T, \{X, T\}, \{W, X, T\}, \{U, X, T\}$, and
$\{W, U, X, T\}$. Note that T is in every conditioning set that *d*-separates Z and Y; that's because
T is the only node in a path that unconditionally *d*-connects Z and Y, so unless it is conditioned
on, Z and Y will always be *d*-connected.

Study questions

Study question 2.4.1

Figure 2.9 below represents a causal graph from which the error terms have been deleted.
Assume that all those errors are mutually independent.

(a) *For each pair of nonadjacent nodes in this graph, find a set of variables that d-separates*
    that pair. What does this list tell us about independencies in the data?
(b) *Repeat question (a) assuming that only variables in the set* $\{Z_3, W, X, Z_1\}$ *can be*
    measured.
(c) *For each pair of nonadjacent nodes in the graph, determine whether they are independent*
    conditional on all other variables.
(d) *For every variable V in the graph, find a minimal set of nodes that renders V independent*
    of all other variables in the graph.
(e) *Suppose we wish to estimate the value of Y from measurements taken on all other variables*
    in the model. Find the smallest set of variables that would yield as good an estimate of Y
    as when we measured all variables.
(f) *Repeat question (e) assuming that we wish to estimate the value of* Z_2.
(g) *Suppose we wish to predict the value of* Z_2 *from measurements of* Z_3. *Would the quality of*
    our prediction improve if we add measurement of W? Explain.

2.5   Model Testing and Causal Search

The preceding sections demonstrate that causal models have *testable implications* in the data
sets they generate. For instance, if we have a graph G that we believe might have generated
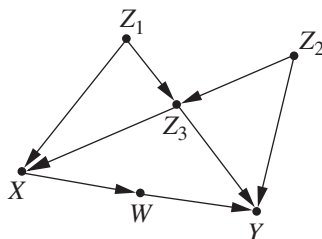
Figure 2.9 A causal graph used in study question 2.4.1, all *U* terms (not shown) are assumed independent

a data set *S*, *d*-separation will tell us which variables in *G* must be independent conditional on which other variables. Conditional independence is something we can test for using a data set. Suppose we list the *d*-separation conditions in *G*, and note that variables *A* and *B* must be independent conditional on *C*. Then, suppose we estimate the probabilities based on *S*, and discover that the data suggests that *A* and *B* are *not* independent conditional on *C*. We can then reject *G* as a possible causal model for *S*.

We can demonstrate it on the causal model of Figure 2.9. Among the many conditional independencies advertised by the model, we find that *W* and Z_1 are independent given *X*, because *X* *d*-separates *W* from Z_1. Now suppose we regress *W* on *X* and Z_1. Namely, we find the line

$$w = r_X x + r_1 z_1$$

that best fits our data. If it turns out that r_1 is not equal to zero, we know that *W* depends on Z_1 given *X* and, consequently, that the model is wrong. [Recall, conditional correlation implies conditional dependence.] Not only do we know that the model is wrong, but we also know where it is wrong; the true model must have a path between *W* and Z_1 that is not *d*-separated by *X*. Finally, this is a theoretical result that holds for all acyclic models with independent errors (Verma and Pearl 1990), and we also know that if every *d*-separation condition in the model matches a conditional independence in the data, then no further test can refute the model. This means that, for any data set whatsoever, one can always find a set of functions *F* for the model and an assignment of probabilities to the *U* terms, so as to generate the data precisely.

There are other methods for testing the fitness of a model. The standard way of evaluating fitness involves a statistical hypothesis testing over the entire model, that is, we evaluate how likely it is for the observed samples to have been generated by the hypothesized model, as opposed to sheer chance. However, since the model is not fully specified, we need to first estimate its parameters before evaluating that likelihood. This can be done (approximately) when we assume a linear and Gaussian model (i.e., all functions in the model are linear and all error terms are normally distributed), because, under such assumptions, the joint distribution (also Gaussian) can be expressed succinctly in terms of the model's parameters, and we can then evaluate the likelihood that the observed samples have been generated by the fully parameterized model (Bollen 1989).

There are, however, a number of issues with this procedure. First, if any parameter cannot be estimated, then the joint distribution cannot be estimated, and the model cannot be tested.

As we shall see in Section 3.8.3, this can occur when some of the error terms are correlated or, equivalently, when some of the variables are unobserved. Second, this procedure tests models globally. If we discover that the model is not a good fit to the data, there is no way for us to determine why that is—which edges should be removed or added to improve the fit. Third, when we test a model globally, the number of variables involved may be large, and if there is measurement noise and/or sampling variation associated with each variable, the test will not be reliable.

d-separation presents several advantages over this global testing method. First, it is nonparametric, meaning that it doesn't rely on the specific functions that connect variables; instead, it uses only the graph of the model in question. Second, it tests models locally, rather than globally. This allows us to identify specific areas, where our hypothesized model is flawed, and to repair them, rather than starting from scratch on a whole new model. It also means that if, for whatever reason, we can't identify the coefficient in one area of the model, we can still get some incomplete information about the rest of the model. (As opposed to the first method, in which if we could not estimate one coefficient, we could not test any part of the model.)

If we had a computer, we could test and reject many possible models in this way, eventually whittling down the set of possible models to only a few whose testable implications do not contradict the dependencies present in the data set. It is a set of models, rather than a single model, because some graphs have indistinguishable implications. A set of graphs with indistinguishable implications is called an *equivalence class*. Two graphs G_1 and G_2 are in the same equivalence class if they share a common skeleton—that is, the same edges, regardless of the direction of those edges—and if they share common *v-structures*, that is, colliders whose parents are not adjacent. Any two graphs that satisfy this criterion have identical sets of d-separation conditions and, therefore, identical sets of testable implications (Verma and Pearl 1990).

The importance of this result is that it allows us to search a data set for the causal models that could have generated it. Thus, not only can we start with a causal model and generate a data set—but we can also start with a data set, and reason back to a causal model. This is enormously useful, since the object of most data-driven research is exactly to find a model that explains the data.

There are other methods of causal search—including some that rely on the kind of global model testing with which we began the section—but a full investigation of them is beyond the scope of this book. Those interested in learning more about search should refer to (Pearl 2000; Pearl and Verma 1991; Rebane and Pearl 1987; Spirtes and Glymour 1991; Spirtes et al. 1993).

Study questions

Study question 2.5.1

(a) *Which of the arrows in Figure 2.9 can be reversed without being detected by any statistical test? [Hint: Use the criterion for equivalence class.]*
(b) *List all graphs that are observationally equivalent to the one in Figure 2.9.*
(c) *List the arrows in Figure 2.9 whose directionality can be determined from nonexperimental data.*

(d) *Write down a regression equation for Y such that, if a certain coefficient in that equation is nonzero, the model of Figure 2.9 is wrong.*

(e) *Repeat question (d) for variable Z_3.*

(f) *Repeat question (e) assuming the X is not measured.*

(g) *How many regression equations of the type described in (d) and (e) are needed to ensure that the model is fully tested, namely, that if it passes all these tests, it cannot be refuted by additional tests of these kind. [Hint: Ensure that you test every vanishing partial regression coefficient that is implied by the product decomposition (1.29).]*

Bibliographical Notes for Chapter 2

The distinction between chains and forks in causal models was made by Simon (1953) and Reichenbach (1956) while the treatment of colliders (or common effect) can be traced back to the English economist Pigou (1911) (see Stigler 1999, pp. 36–41). In epidemiology, colliders came to be associated with "Selection bias" or "Berkson paradox" (Berkson 1946) while in artificial intelligence it came to be known as the "explaining away effect" (Kim and Pearl 1983). The rule of d-separation for determining conditional independence by graphs (Definition 2.4.1) was introduced in Pearl (1986) and formally proved in Verma and Pearl (1988) using the theory of graphoids (Pearl and Paz 1987). Gentle introductions to d-separation are available in Hayduk et al. (2003), Glymour and Greenland (2008), and Pearl (2000, pp. 335–337). Algorithms and software for detecting d-separation, as well as finding minimal separating sets are described in Tian et al. (1998), Kyono (2010), and Textor et al. (2011). The advantages of local over global model testing, are discussed in Pearl (2000, pp. 144–145) and further elaborated in Chen and Pearl (2014). Recent applications of d-separation include extrapolation across populations (Pearl and Bareinboim 2014), recovering from sampling selection bias (Bareinboim et al. 2014), and handling missing data (Mohan et al. 2013).

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 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

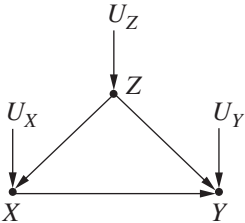$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{3.1}$$

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

which is known as the "causal effect difference," or "average causal effect" (ACE). In general, however, if X and Y can each take on more than one value, we would wish to predict the general causal effect $P(Y = y|do(X = x))$, where x and y are any two values that X and Y can take on. For example, x may be the dosage of the drug and y the patient's blood pressure.

We know from first principles that causal effects cannot be estimated from the data set itself without a causal story. That was the lesson of Simpson's paradox: The data itself was not sufficient even for determining whether the effect of the drug was positive or negative. But with the aid of the graph in Figure 3.3, we can compute the magnitude of the causal effect from the data. To do so, we simulate the intervention in the form of a graph surgery (Figure 3.4) just as we did in the ice cream example. The causal effect $P(Y = y|do(X = x))$ is equal to the conditional probability $P_m(Y = y|X = x)$ that prevails in the *manipulated* model of Figure 3.4. (This, of course, also resolves the question of whether the correct answer lies in the aggregated or the Z-specific table—when we determine the answer through an intervention, there's only one table to contend with.)
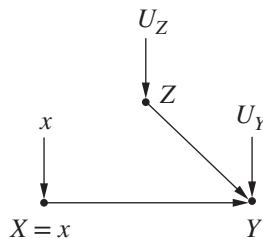

Figure 3.4 A modified graphical model representing an intervention on the model in Figure 3.3 that sets drug usage in the population, and results in the manipulated probability P_m

The key to computing the causal effect lies in the observation that P_m, the manipulated probability, shares two essential properties with P (the original probability function that prevails in the preintervention model of Figure 3.3). First, the marginal probability $P(Z = z)$ is invariant under the intervention, because the process determining Z is not affected by removing the arrow from Z to X. In our example, this means that the proportions of males and females remain the same, before and after the intervention. Second, the conditional probability $P(Y = y|Z = z, X = x)$ is invariant, because the process by which Y responds to X and Z, $Y = f(x, z, u_Y)$, remains the same, regardless of whether X changes spontaneously or by deliberate manipulation. We can therefore write two equations of invariance:

$$P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x) \quad \text{and} \quad P_m(Z = z) = P(Z = z)$$

We can also use the fact that Z and X are d-separated in the modified model and are, therefore, independent under the intervention distribution. This tells us that $P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$, the last equality following from above. Putting these considerations together, we have

$$P(Y = y|do(X = x))$$

$$= P_m(Y = y|X = x) \qquad \text{(by definition)} \tag{3.2}$$

$$= \sum_z P_m(Y = y | X = x, Z = z) P_m(Z = z | X = x) \tag{3.3}$$

$$= \sum_z P_m(Y = y | X = x, Z = z) P_m(Z = z) \tag{3.4}$$

Equation (3.3) is obtained ~~from Bayes' rule~~ using the Law of Total Probability by conditioning on and summing over all values of $Z = z$ (as in Eq. (1.9)), while (Eq.(3.4) makes use of the independence of Z and X in the modified model.

Finally, using the invariance relations, we obtain a formula for the causal effect, in terms of preintervention probabilities:

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, Z = z) P(Z = z) \tag{3.5}$$

Equation (3.5) is called the *adjustment formula*, and as you can see, it computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred to as "adjusting for Z" or "controlling for Z."

This final expression—the right-hand side of Eq. (3.5)—can be estimated directly from the data, since it consists only of conditional probabilities, each of which can be computed by the filtering procedure described in Chapter 1. Note also that no adjustment is needed in a randomized controlled experiment since, in such a setting, the data are generated by a model which already possesses the structure of Figure 3.4, hence, $P_m = P$ regardless of any factors Z that affect Y. Our derivation of the adjustment formula (3.5) constitutes therefore a formal proof that randomization gives us the quantity we seek to estimate, namely $P(Y = y | do(X = x))$. In practice, investigators use adjustments in randomized experiments as well, for the purpose of minimizing sampling variations (Cox 1958).

To demonstrate the working of the adjustment formula, let us apply it numerically to Simpson's story, with $X = 1$ standing for the patient taking the drug, $Z = 1$ standing for the patient being male, and $Y = 1$ standing for the patient recovering. We have

$$P(Y = 1 | do(X = 1)) = P(Y = 1 | X = 1, Z = 1) P(Z = 1) + P(Y = 1 | X = 1, Z = 0) P(Z = 0)$$

Substituting the figures given in Table 1.1 we obtain

$$P(Y = 1 | do(X = 1)) = \frac{0.93(87 + 270)}{700} + \frac{0.73(263 + 80)}{700} = 0.832$$

while, similarly,

$$P(Y = 1 | do(X = 0)) = \frac{0.87(87 + 270)}{700} + \frac{0.69(263 + 80)}{700} = 0.7818$$

Thus, comparing the effect of drug-taking ($X = 1$) to the effect of nontaking ($X = 0$), we obtain

$$ACE = P(Y = 1 | do(X = 1)) - P(Y = 1 | do(X = 0)) = 0.832 - 0.7818 = 0.0502$$

giving a clear positive advantage to drug-taking. A more informal interpretation of ACE here is that it is simply the difference in the fraction of the population that would recover if everyone took the drug compared to when no one takes the drug.

We see that the adjustment formula instructs us to condition on gender, find the benefit of the drug separately for males and females, and only then average the result using the percentage of males and females in the population. It also thus instructs us to ignore the aggregated

these parents that we neutralize when we fix X by external manipulation. Denoting the parents of X by $PA(X)$, we can therefore write a general adjustment formula and summarize it in a rule:

Rule 1 (The Causal Effect Rule) *Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by*

$$P(Y = y | do(X = x)) = \sum_z P(Y = y | X = x, PA = z) P(PA = z) \qquad (3.6)$$

where z ranges over all the combinations of values that the variables in PA can take.

If we multiply and divide the summand in (3.6) by the probability $P(X = x | PA = z)$, we get a more convenient form:

$$P(y | do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x | PA = z)} \qquad (3.7)$$

which explicitly displays the role played by the parents of X in predicting the results of interventions. The factor $P(X = x | PA = z)$ is known as the "propensity score" and the advantages of expressing $P(y | do(x))$ in this form will be discussed in Section 3.5.

We can appreciate now what role the causal graph plays in resolving Simpson's paradox, and, more generally, what aspects of the graph allow us to predict causal effects from purely statistical data. We need the graph in order to determine the identity of X's parents—the set of factors that, under nonexperimental conditions, would be sufficient for determining the value of X, or the probability of that value.

This result alone is astounding; using graphs and their underlying assumptions, we were able to identify causal relationships in purely observational data. But, from this discussion, readers may be tempted to conclude that the role of graphs is fairly limited; once we identify the parents of X, the rest of the graph can be discarded, and the causal effect can be evaluated mechanically from the adjustment formula. The next section shows that things may not be so simple. In most practical cases, the set of X's parents will contain unobserved variables that would prevent us from calculating the conditional probabilities in the adjustment formula. Luckily, as we will see in future sections, we can adjust for other variables in the model to substitute for the unmeasured elements of $PA(X)$.

Study questions

Study questions 3.2.1

Referring to Study question 1.5.2 (Figure 1.10) and the parameters listed therein,

(a) *Compute $P(y | do(x))$ for all values of x and y, by simulating the intervention do(x) on the model.*
(b) *Compute $P(y | do(x))$ for all values of x and y, using the adjustment formula (3.5)*
(c) *Compute the ACE*

$$ACE = P(y_1 | do(x_1)) - P(y_1 | do(x_0))$$

and compare it to the Risk Difference

$$RD = P(y_1|x_1) - P(y_1|x_0)$$

What is the difference between ACE and the RD? What values of the parameters would minimize the difference?

(d) *Find a combination of parameters that exhibit Simpson's reversal (as in Study question 1.5.2(c)) and show explicitly that the overall causal effect of the drug is obtained from the* ~~desegregated~~ disaggregated *data.*

3.2.2 Multiple Interventions and the Truncated Product Rule

In deriving the adjustment formula, we assumed an intervention on a single variable, X, whose parents were disconnected, so as to simulate the absence of their influence after intervention. However, social and medical policies occasionally involve multiple interventions, such as those that dictate the value of several variables simultaneously, or those that control a variable over time. To represent multiple interventions, it is convenient to resort to the product decomposition that a graphical model imposes on joint distributions, as we have discussed in Section 1.5.2. According to the Rule of Product Decomposition, the preintervention distribution in the model of Figure 3.3 is given by the product

$$P(x, y, z) = P(z)P(x|z)P(y|x, z) \tag{3.8}$$

whereas the postintervention distribution, governed by the model of Figure 3.4 is given by the product

$$P(z, y|do(x)) = P_m(z)P_m(y|x, z) = P(z)P(y|x, z) \tag{3.9}$$

with the factor $P(x|z)$ purged from the product, since X becomes parentless as it is fixed at $X = x$. This coincides with the adjustment formula, because to evaluate $P(y|do(x))$ we need to marginalize (or sum) over z, which gives

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

in agreement with (3.5).

This consideration also allows us to generalize the adjustment formula to multiple interventions, that is, interventions that fix the values of a set of variables X to constants. We simply write down the product decomposition of the preintervention distribution, and strike out all factors that correspond to variables in the intervention set X. Formally, we write

$$P(x_1, x_2, \dots, x_n|do(x)) = \prod_i P(x_i|pa_i) \qquad \text{for all } i \text{ with } X_i \text{ not in } X.$$

This came to be known as the *truncated product formula* or *g-formula*. To illustrate, assume that we intervene on the model of Figure 2.9 and set X to x and Z_3 to z_3. The postintervention distribution of the other variables in the model will be

$$P(z_1, z_2, w, y|do(X = x, Z_3 = z_3)) = P(z_1)P(z_2)P(w|x)P(y|w, z_3, z_2)$$

where we have deleted the factors $P(x|z_1, z_3)$ and $P(z_3|z_1, z_2)$ from the product.

It is interesting to note that combining (3.8) and (3.9), we get a simple relation between the pre- and postintervention distributions:

$$P(z, y|do(x)) = \frac{P(x, y, z)}{P(x|z)} \tag{3.10}$$

It tells us that the conditional probability $P(x|z)$ is all we need to know in order to predict the effect of an intervention $do(x)$ from nonexperimental data governed by the distribution $P(x, y, z)$.

3.3 The Backdoor Criterion

In the previous section, we came to the conclusion that we should adjust for a variable's parents, when trying to determine its effect on another variable. But often, we know, or believe, that the variables have unmeasured parents that, though represented in the graph, may be inaccessible for measurement. In those cases, we need to find an alternative set of variables to adjust for.

This dilemma unlocks a deeper statistical question: Under what conditions does a causal story permit us to compute the causal effect of one variable on another, from data obtained by passive observations, with no interventions? Since we have decided to represent causal stories with graphs, the question becomes a graph-theoretical problem: Under what conditions is the structure of the causal graph sufficient for computing a causal effect from a given data set?

The answer to that question is long enough—and important enough—that we will spend the rest of the chapter addressing it. But one of the most important tools we use to determine whether we can compute a causal effect is a simple test called the *backdoor criterion*. Using it, we can determine whether, for any two variables X and Y in a causal model represented by a DAG, which set of variables Z in that model should be conditioned on when searching for the causal relationship between X and Y.

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

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

just as when we adjust for $PA(X)$. (Note that $PA(X)$ always satisfies the backdoor criterion.)

The logic behind the backdoor criterion is fairly straightforward. In general, we would like to condition on a set of nodes Z such that

1. We block all spurious paths between X and Y.
2. We leave all directed paths from X to Y unperturbed.
3. We create no new spurious paths.

When trying to find the causal effect of X on Y, we want the nodes we condition on to block any "backdoor" path in which one end has an arrow into X, because such paths may make X and Y dependent, but are obviously not transmitting causal influences from X, and if we do not block them, they will confound the effect that X has on Y. We condition on backdoor paths so as to fulfill our first requirement. However, we don't want to condition on any nodes that are descendants of X. Descendants of X would be affected by an intervention on X and might themselves affect Y; conditioning on them would block those pathways. Therefore, we don't condition on descendants of X so as to fulfill our second requirement. Finally, to comply with the third requirement, we should refrain from conditioning on any collider that would unblock a new path between X and Y. The requirement of excluding descendants of X also protects us from conditioning on children of intermediate nodes between X and Y (e.g., the collision node W in Figure 2.4.) Such conditioning would distort the passage of causal association between X and Y, similar to the way conditioning on their parents would.

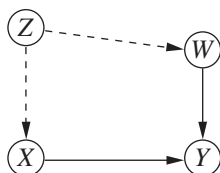To see what this means in practice, let's look at a concrete example, shown in Figure 3.6.

Figure 3.6 A graphical model representing the relationship between a new drug (X), recovery (Y), weight (W), and an unmeasured variable Z (socioeconomic status)

Here we are trying to gauge the effect of a drug (X) on recovery (Y). We have also measured weight (W), which has an effect on recovery. Further, we know that socioeconomic status (Z) affects both weight and the choice to receive treatment—but the study we are consulting did not record socioeconomic status.

Instead, we search for an observed variable that fits the backdoor criterion from X to Y. A brief examination of the graph shows that W, which is not a descendant of X, also blocks the backdoor path $X \leftarrow Z \rightarrow W \rightarrow Y$. Therefore, W meets the backdoor criterion. So long as the causal story conforms to the graph in Figure 3.6, adjusting for W will give us the causal effect of X on Y. Using the adjustment formula, we find

$$P(Y = y | do(X = x)) = \sum_w P(Y = y | X = x, W = w) P(W = w)$$

This sum can be estimated from our observational data, so long as W is observed.

With the help of the backdoor criterion, you can easily and algorithmically come to a conclusion about a pressing policy concern, even in complicated graphs. Consider the model in Figure 2.8, and assume again that we wish to evaluate the effect of X on Y. What variables should we condition on to obtain the correct effect? The question boils down to finding a set of variables that satisfy the backdoor criterion, but since there are no backdoor paths from X to Y, the answer is trivial: The empty set satisfies the criterion, hence no adjustment is needed. The answer is

$$P(y | do(x)) = P(y | x)$$

Suppose, however, that we were to adjust for W. Would we get the correct result for the effect of X on Y? Since W is a collider, conditioning on W would open the path $X \rightarrow W \leftarrow Z$

$T \rightarrow Y$. 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? In Figure 2.8, 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.
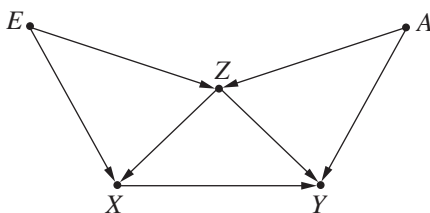

Figure 3.7 A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y

(c) Write the adjustment formula for the effect of the drug on recovery.
(d) Repeat questions (a)–(c) assuming that the nurse gave lollipops a day after the study, still preferring patients who received treatment over those who received placebo.

3.4 The Front-Door Criterion

The backdoor criterion provides us with a simple method of identifying sets of covariates that should be adjusted for when we seek to estimate causal effects from nonexperimental data. It does not, however, exhaust *all* ways of estimating such effects. The *do*-operator can be applied to graphical patterns that do not satisfy the backdoor criterion to identify effects that on first sight seem to be beyond one's reach. One such pattern, called front-door, is discussed in this section.

Consider the century-old debate on the relation between smoking and lung cancer. In the years preceding 1970, the tobacco industry has managed to prevent antismoking legislation by promoting the theory that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype that also induces an inborn craving for nicotine.
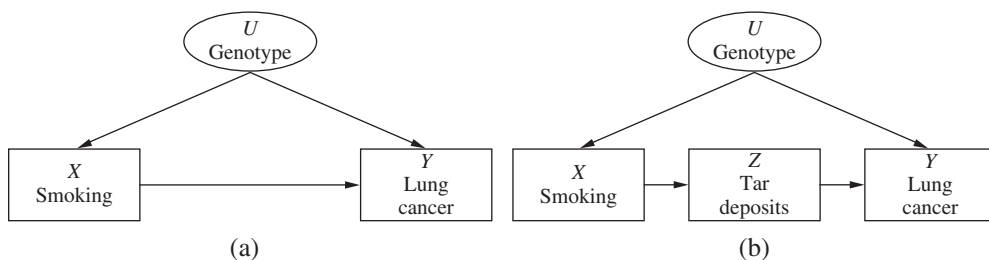

Figure 3.10 A graphical model representing the relationships between smoking (X) and lung cancer (Y), with unobserved confounder (U) and a mediating variable Z

A graph depicting this example is shown in Figure 3.10(a) This graph does not satisfy the backdoor condition because the variable U is unobserved and hence cannot be used to block the backdoor path from X to Y. The causal effect of smoking on lung cancer is not identifiable in this model; one can never ascertain which portion of the observed correlation between X and Y is spurious, attributable to their common effect, U, and what portion is genuinely causative. (We note, however, that even in these circumstances, much compelling work has been done to quantify how strong the (unobserved) associates between both U and X, and U and Y, must be in order to entirely explain the observed association between X and Y.)

However, we can go much further by considering the model in Figure 3.10(b), where an additional measurement is available: the amount of tar deposits in patients lungs. This model does not satisfy the backdoor criterion, because there is still no variable capable of blocking the spurious path $X \leftarrow U \rightarrow Y$. We see, however, that the causal effect $P(Y = y|do(X = x))$ is nevertheless identifiable in this model, through two consecutive applications of the backdoor criterion.

How can the intermediate variable Z help us to assess the effect of X on Y? The answer is not at all trivial: as the following quantitative example shows, it may lead to heated debate.

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, ~~random,~~ 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%)

It appears that tar deposits have a harmful effect in both groups; in smokers it increases cancer rates from 10% to 15%, and in nonsmokers it increases cancer rates from 90% to 95%. Thus, regardless of whether I have a natural craving for nicotine, I should avoid the harmful effect of tar deposits, and no-smoking offers very effective means of avoiding them.

The graph of Figure 3.10(b) enables us to decide between these two groups of statisticians. First, we note that the effect of X on Z is identifiable, since there is no backdoor path from X to Z. Thus, we can immediately write

$$P(Z = z|do(X = x)) = P(Z = z|X = x) \tag{3.12}$$

Next we note that the effect of Z on Y is also identifiable, since the backdoor path from Z to Y, namely $Z \leftarrow X \leftarrow U \rightarrow Y$, can be blocked by conditioning on X. Thus, we can write

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x)P(X = x) \tag{3.13}$$

Both (3.12) and (3.13) are obtained through the adjustment formula, the first by conditioning on the null set, and the second by adjusting for X.

We are now going to chain together the two partial effects to obtain the overall effect of X on Y. The reasoning goes as follows: If nature chooses to assign Z the value z, then the probability of Y would be $P(Y = y|do(Z = z))$. But the probability that nature would choose to do that, given that we choose to set X at x, is $P(Z = z|do(X = x))$. Therefore, summing over all states z of Z, we have

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))P(Z = z|do(X = x)) \tag{3.14}$$

The terms on the right-hand side of (3.14) were evaluated in (3.12) and (3.13), and we can substitute them to obtain a *do*-free expression for $P(Y = y|do(X = x))$. We also distinguish between the x that appears in (3.12) and the one that appears in (3.13), the latter of which is merely an index of summation and might as well be denoted x'. The final expression we have is

$$P(Y = y|do(X = x)) =$$
$$\sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x) \tag{3.15}$$

Equation (3.15) is known as the *front-door formula*.

Applying this formula to the data in Table 3.1, we see that the tobacco industry was wrong; tar deposits have a harmful effect in that they make lung cancer more likely and smoking, by increasing tar deposits, increases the chances of causing this harm.

The data in Table 3.1 are obviously unrealistic and were deliberately crafted so as to surprise readers with counterintuitive conclusions that may emerge from naive analysis of observational data. In reality, we would expect observational studies to show positive correlation between smoking and lung cancer. The estimand of (3.15) could then be used for confirming and quantifying the harmful effect of smoking on cancer.

The preceding analysis can be generalized to structures, where multiple paths lead from X to Y.

Definition 3.4.1 (Front-Door) *A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if*

1. *Z intercepts all directed paths from X to Y.*
2. *There is no ~~unblocked~~ backdoor path from X to Z.*
3. *All backdoor paths from Z to Y are blocked by X.*

Theorem 3.4.1 (Front-Door Adjustment) *If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{3.16}$$

The conditions stated in Definition 3.4.1 are overly conservative; some of the backdoor paths excluded by conditions (2) and (3) can actually be allowed provided they are blocked by some variables. There is a powerful symbolic machinery, called the *do-calculus*, that allows analysis of such intricate structures. In fact, the *do*-calculus uncovers *all* causal effects that can be identified from a given graph. Unfortunately, it is beyond the scope of this book (see ~~Pearl 2009 and Shpitser and Pearl 2008~~ for details). But the combination of the adjustment formula, the backdoor criterion, and the front-door criterion covers numerous scenarios. It proves the enormous, even revelatory, power that causal graphs have in not merely representing, but actually discovering causal information.

Study questions

Study question 3.4.1

Assume that in Figure 3.8, only X, Y, and one additional variable can be measured. Which variable would allow the identification of the effect of X on Y? What would that effect be?

Study question 3.4.2

I went to a pharmacy to buy a certain drug, and I found that it was available in two different bottles: one priced at $1, the other at $10. I asked the druggist, "What's the difference?" and he told me, "The $10 bottle is fresh, whereas the $1 bottle one has been on the shelf for 3 years. But, you know, data shows that the percentage of recovery is much higher among those who bought the cheap stuff. Amazing isn't it?" I asked if the aged drug was ever tested. He said, "Yes, and this is even more amazing; 95% of the aged drug and only 5% of the fresh drug has lost the active ingredient, yet the percentage of recovery among those who got bad bottles, with none of the active ingredient, is still much higher than among those who got good bottles, with the active ingredient."

Before ordering a cheap bottle, it occurred to me to have a good look at the data. The data were, for each previous customer, the type of bottle purchased (aged or fresh), the concentration of the active ingredient in the bottle (high or low), and whether the customer recovered from the illness. The data perfectly confirmed the druggist's story. However, after making some additional calculations, I decided to buy the expensive bottle after all; even without testing its

Tian and Pearl 2002, Shpitser and Pearl 2008, Pearl 2009, and Bareinboim and Pearl 2012

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 may 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

$$P(Y = y | do(X = g(Z)))$$

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

$$= \sum_z P(Y = y | do(X = g(z)), Z = z) P(Z = z) \qquad (3.17)$$

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

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

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.

(b) *Identify a set of four variables that need to be measured in order to estimate the z-specific effect of X on Y, and find an expression for the size of that effect.*

(c) *Using your answer to part (b), determine the expected value of Y under a Z-dependent strategy, where X is set to 0 when Z is smaller or equal to 2 and X is set to 1 when Z is larger than 2. (Assume Z takes on integer values from 1 to 5.)*

3.6   Inverse Probability Weighing

By now, the astute reader may have noticed a problem with our intervention procedures. The backdoor and front-door criteria tell us whether it is possible to predict the results of hypothetical interventions from data obtained in an observational study. Moreover, they tell us that we can make this prediction without simulating the intervention and without even thinking about it. All we need to do is identify a set Z of covariates satisfying one of the criteria, plug this set into the adjustment formula, and we're done: the resulting expression is guaranteed to provide a valid prediction of how the intervention will affect the outcome.

This is lovely in theory, but in practice, adjusting for Z may prove problematic. It entails looking at each value or combination of values of Z separately, estimating the conditional probability of Y given X in that stratum and then averaging the results. As the number of strata increases, adjusting for Z will encounter both computational and estimational difficulties. Since the set Z can be comprised of dozens of variables, each spanning dozens of discrete values, the summation required by the adjustment formula may be formidable, and the number of data samples falling within each $Z = z$ cell may be too small to provide reliable estimates of the conditional probabilities involved.

All of our work in this chapter has not been for naught, however. The adjustment procedure is straightforward, and, therefore, easy to use in the explanation of intervention criteria. But there is another, more subtle procedure that overcomes the practical difficulties of adjustment.

In this section, we discuss one way of circumventing this problem, provided only that we can obtain a reliable estimate of the function $g(x, z) = P(X = x | Z = z)$, often called the "propensity score," for each x and z. Such an estimate can be obtained by fitting the parameters of a flexible function $g(x, z)$ to the data at hand, in much the same way that we fitted the coefficients of a linear regression function, so as to minimize the mean square error with respect to a set of samples (Figure 1.4). The method used will depend on the nature of the random variable X, whether it is continuous, discrete, binary, or a caveat, for example.

Assuming that the function $P(X = x | Z = z)$ is available to us, we can use it to generate artificial samples that act as though they were drawn from the postintervention probability P_m, rather than $P(x, y, z)$. Once we obtain such fictitious samples, we can evaluate $P(Y = y | do(x))$ by simply counting the frequency of the event $Y = y$, for each stratum $X = x$ in the sample. In this way, we skip the labor associated with summing over all strata $Z = z$; we essentially let nature do the summation for us.

The idea of estimating probabilities using fictitious samples is not new to us; it was used all along, though implicitly, whenever we estimated conditional probabilities from finite samples.

In Chapter 1, we characterized conditioning as a process of filtering—that is, ignoring all cases for which the condition $X = x$ does not hold, and normalizing the surviving cases, so that their total probabilities would add up to one. The net result of this operation is that the probability of each surviving case is boosted by a factor $1/P(X = x)$. This can be seen directly

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.")

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
($X = yes$) in the population of Table 3.3

X	Y	Z	% of population
Yes	Yes	Male	~~0.232~~ 0.231
Yes	Yes	Female	~~0.547~~ 0.549
Yes	No	Male	~~0.02~~ 0.017
Yes	No	Female	~~0.202~~ 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 "$X = Yes$," we get the data set shown in Table 3.4, which was formed in two steps. First, all rows with $X = No$ 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 $1/P(X = yes)$, and $P(X = yes)$ in our example, is the combined weight of the first four rows of Table 3.3, which amounts to

$$P(X = yes) = 0.116 + 0.274 + 0.01 + 0.101 = 0.501$$

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 $1/0.501 = 2.00$.

Let us now examine the population created by the $do(X = yes)$ 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 $P(X = yes|Z = z)$ for each z, which, according to Table 3.3, is given by

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

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

Multiplying the gender-matching rows by $1/0.247$ 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:

$$P(Y = yes|do(X = yes)) = 0.476 + 0.357 = 0.833$$

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

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

the *controlled direct effect* (CDE) on Y of changing the value of X from x to x' is defined as

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) \qquad (3.18)$$

The obvious advantage of this definition over the one based on conditioning is its generality; it captures the intent of "keeping Z constant" even in cases where the $Z \rightarrow Y$ relationship is confounded (the same goes for the $X \rightarrow Z$ and $X \rightarrow Y$ relationships). Practically, this definition assures us that in any case where the intervened probabilities are identifiable from the observed probabilities, we can estimate the direct effect of X on Y. Note that the direct effect may differ for different values of Z; for instance, it may be that hiring practices discriminate against women in jobs with high qualification requirements, but they discriminate against men in jobs with low qualifications. Therefore, to get the full picture of the direct effect, we'll have to perform the calculation for every relevant value z of Z. (In linear models, this will not be necessary; for more information, see Section 3.8.)
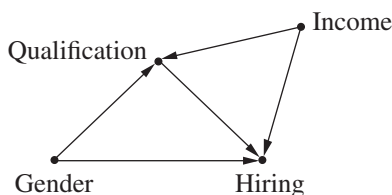

Figure 3.12   A graphical model ~~repr~~ showing qualification (Z) as a mediator between gender (X) and hiring (Y), ~~with socioeconomic status as a media~~ and income (I) as a confounder between qualification and hiring.

How do we estimate the direct effect when its expression contains two *do*-operators? The technique is more or less the same as the one employed in Section 3.2, where we dealt with a single *do*-operator by adjustment. In our example of Figure 3.12, we first notice that there is no backdoor path from X to Y in the model, hence we can replace $do(x)$ with simply conditioning on x (this essentially amounts to adjusting for all confounders). This results in

$$P(Y = y|X = x, do(Z = z)) - P(Y = y|X = x', do(Z = z))$$

Next, we attempt to remove the $do(z)$ term and notice that two backdoor paths exist from Z to Y, one through X and one through I. The first is blocked (since X is conditioned on) and the second can be blocked if we adjust for I. This gives

$$\sum_i [P(Y = y|X = x, Z = z, I = i) - P(Y = y|X = x', Z = z, I = i)]P(I = i)$$

The last formula is *do*-free, which means it can be estimated from nonexperimental data.

In general, the CDE of X on Y, mediated by Z, is identifiable if the following two properties hold:

1. There exists a set S_1 of variables that blocks all backdoor paths from Z to Y.
2. There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z.

enormous simplification of the procedure needed for causal analysis. We are all familiar with the bell-shaped curve that characterizes the normal distribution of one variable. The reason it is so popular in statistics is that it occurs so frequently in nature whenever a phenomenon is a byproduct of many noisy microprocesses that add up to produce macroscopic measurements such as height, weight, income, or mortality. Our interest in the normal distribution, however, stems primarily from the way several normally distributed variables combine to shape their joint distribution. The assumption of normality gives rise to four properties that are of enormous use when working with linear systems:

1. Efficient representation
2. Substitutability of expectations for probabilities
3. Linearity of expectations
4. Invariance of regression coefficients.

Starting with two normal variables, X and Y, we know that their joint density forms a three-dimensional cusp (like a mountain rising above the X–Y plane) and that the planes of equal height on that cusp are ellipses like those shown in Figure 1.2. Each such ellipse is characterized by five parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y,$ and ρ_{XY}, as defined in Sections 1.3.8 and 1.3.9. The parameters μ_X and μ_Y specify the location (or the center of gravity) of the ellipse in the X–Y plane, the ~~variances~~ *standard deviations* σ_X and σ_Y specify the spread of the ellipse along the X and Y dimensions, respectively, and the correlation coefficient ρ_{XY} specifies its orientation. In three dimensions, the best way to depict the joint distribution is to imagine an oval football suspended in the X–Y–Z space (Figure 1.2); every plane of constant Z would then cut the football in a two-dimensional ellipse like the ones shown in Figure 1.1.

As we go to higher dimensions, and consider a set of N normally distributed variables $X_1, X_2, \ldots, X_N$, we need not concern ourselves with additional parameters; it is sufficient to specify those that characterize the $N(N-1)/2$ pairs of variables, (X_i, X_j). In other words, the joint density of $(X_1, X_2, \ldots, X_N)$ is fully specified once we specify the bivariate density of (X_i, X_j), with i and j $(i \neq j)$ ranging from 1 to N. This is an enormously useful property, as it offers an extremely parsimonious way of specifying the N-variable joint distribution. Moreover, since the joint distribution of each pair is specified by five parameters, we conclude that the joint distribution requires at most $5 \times N(N-1)/2$ parameters (means, variances, and covariances), each defined by expectation. In fact, the total number of parameters is even smaller than this, namely $2N + N(N-1)/2$; the first term gives the number of mean and variance parameters, and the second the number of correlations.

This brings us to another useful feature of multivariate normal distributions: they are fully defined by expectations, so we need not concern ourselves with probability tables as we did when dealing with discrete variables. Conditional probabilities can be expressed as conditional expectations, and notions such as conditional independence that define the structure of graphical models can be expressed in terms of equality relationships among conditional expectations. For instance, to express the conditional independence of Y and X, given Z,

$$P(Y|X, Z) = P(Y|Z)$$

we can write

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

(where Z is a set of variables).

This feature of normal systems gives us an incredibly useful ability: Substituting expectations for probabilities allows us to use regression (a predictive method) to determine causal information. The next useful feature of normal distributions is their linearity: every conditional expectation $E[Y|X_1, X_2, \ldots, X_n]$ is given by a linear combination of the conditioning variables. Formally,

$$E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = r_0 + r_1 x_1 + r_2 x_2 + \cdots + r_n x_n$$

where each of the slopes $r_1, r_2, \ldots, r_n$ is a partial regression coefficient ~~of~~ as defined in Sections 1.3.10 and 1.3.11.

The magnitudes of these slopes do not depend on the values $x_1, x_2, \ldots, x_n$ of the conditioning variables, called *regressors*; they depend only on which variables are chosen as regressors. In other words, the sensitivity of Y to the measurement $X_i = x_i$ does not depend on the measured values of the other variables in the regression; it depends only on which variables we choose to measure. It doesn't matter whether $X_i = 1$, $X_i = 2$, or $X_i = 312.3$; as long as we regress Y on ~~X_i, the coefficient r_j of X_j will remain the same.~~ $X_1, X_2, \ldots, X_n$ all slopes will remain the same.

This unique and useful feature of normal distributions is illustrated in Figures 1.1 and 1.2 of Chapter 1. Figure 1.1 shows that regardless of what level of age we choose, the slope of Y on X at that level is the same. If, however, we do not hold age constant (i.e., we do not regress on it), the slope becomes vastly different, as is shown in Figure 1.2.

The linearity assumption also permits us to fully specify the functions in the model by annotating the causal graph with a *path coefficient* (or structural coefficient) along each edge. The path coefficient β along the edge $X \rightarrow Y$ quantifies the contribution of X in the function that defines Y in the model. For instance, if the function defines $Y = 3X + U$, the path coefficient of $X \rightarrow Y$ will be 3. The path coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are fundamentally different from the regression coefficients $r_1, r_2, \ldots, r_n$ that we discussed in Section 1.3. The former are "structural" or "causal," whereas the latter are statistical. The difference is explained in the next section.

Many of the regression methods we discuss are far more general, applying in situations where the variables $X_1, \ldots, X_k$ follow distribution far from multivariate Normal; for example, when some of the X_i's are categorical or even binary. Such generalizations also therefore allow the conditional mean $E(Y|X_1 = x_1, \ldots, X_k = x_k)$ to include nonlinear combinations of the X_i's, including such terms as $X_1 X_2$, for example, to allow for effect modification, or interaction. Since we are conditioning on the values of the X_i's, it is usually not necessary to enforce a distributional assumption for such variables. Nevertheless, the full multivariate Normal scenario provides considerable insight into structural causal models.

3.8.1 Structural versus Regression Coefficients

As we are now about to deal with linear models, and thus, as a matter of course, with regression-like equations, it is of paramount importance to define the difference between regression equations and the structural equations we have used in SCMs throughout the book. A regression equation is descriptive; it makes no assumptions about causation. When we write $y = r_1 x + r_2 z + \epsilon$, as a regression equation, we are not saying that X and Z cause Y. We merely confess our need to know which values of r_1 and r_2 would make the equation $y = r_1 x + r_2 z$

the best linear approximation to the data, or, equivalently, the best linear approximation of $E(y|x, z)$.

Because of this fundamental difference between structural and regression equations, some books distinguish them by writing an arrow, instead of equality sign, in structural equations, and some distinguish the coefficients by using a different font. We distinguish them by denoting structural coefficients as α, β, and so on, and regression coefficients as r_1, r_2, and so on. In addition, we distinguish between the stochastic "error terms" that appear in these equations. Errors in regression equations are denoted ϵ_1, ϵ_2, and so on, as in Eq. (1.24), and those in structural equations by U_1, U_2, and so on, as in SCM 1.5.2. The former denote the residual errors in observation, after fitting the equation $y = r_1 x + r_2 z$ to data, whereas the latter represent latent factors (sometimes called "disturbances" or "omitted variables") that influence Y and are not themselves affected by X. The former are human-made (due to imperfect fitting); the latter are nature-made.

Though they are not causally binding themselves, regression equations are of significant use in the study of causality as it pertains to linear systems. Consider: In Section 3.2, we were able to express the effects of interventions in terms of conditional probabilities, as, for example, in the adjustment formula of Eq. (3.5). In linear systems, the role of conditional probabilities will be taken over by regression coefficients, since these coefficients represent the dependencies induced by the model and, in addition, they are easily estimable using least square analyses. Similarly, whereas the testable implications of nonparametric models are expressed in the form of conditional independencies, these independencies are signified in linear models by vanishing regression coefficients, like those discussed in Section 1.3.11. Specifically, given the regression equation

$$y = r_0 + r_1 x_1 + r_2 x_2 + \cdots + r_n x_n + \epsilon$$

if $r_i = 0$, then Y is independent of X_i conditional on all the other regression variables.

3.8.2 *The Causal Interpretation of Structural Coefficients*

In a linear system, every path coefficient stands for the direct effect of the independent variable, X, on the dependent variable, Y. To see why this is so, we refer to the interventional definition of direct effect given in Section 3.7 (Eq. (3.18)), which calls for computing the change in Y as X increases by one unit whereas all other parents of Y are held constant. When we apply this definition to any linear system, regardless of whether the disturbances are correlated or not, the result will be the path coefficient on the arrow $X \rightarrow Y$.

Consider, for example, the model in Figure 3.13, and assume we wish to estimate the direct effect of Z on Y. The structural equations in the fully specified model read:

$$X = U_X$$
$$Z = aX + U_Z$$
$$W = bX + cZ + U_W$$
$$Y = dZ + eW + U_Y$$

is the sum of the products of the coefficients on the two nonbackdoor paths from Z to Y. This will be the case in all linear models; algebra demands it. Moreover, the sum of product rule will be valid regardless of the distributions of the U variables and regardless of whether they are dependent or independent.

3.8.3   *Identifying Structural Coefficients and Causal Effect*

Thus far, we have expressed the total and direct effects in terms of path coefficients, assuming that the latter are either known to us a priori or estimated from interventional experiments. We now tackle a much harder problem; estimating total and direct effects from nonexperimental data. This problem is known as "identifiability" and, mathematically, it amounts to expressing the path coefficients associated with the total and direct effects in terms of the covariances σ_{XY} or regression coefficients $R_{YX \cdot Z}$, where X and Y are any two variables in the model, and Z a set of variables in the model (Eqs. (1.27) and (1.28) and Section 1.3.11).

In many cases, however, it turns out that to identify direct and total effects, we do not need to identify each and every structural parameter in the model. Let us first demonstrate with the total effect, τ. The backdoor criterion gives us the set Z of variables we need to adjust for in order to determine the causal effect of X on Y. How, though, do we make use of the criterion to determine effects in a linear system? In principle, once we obtain the set, Z, we can estimate the conditional expectation of Y given X and Z and, then, averaging over Z, we can use the resultant dependence between Y and X to measure the effect of X on Y. We need only translate this procedure to the language of regression.

The translation is rather simple. First, we find a set of covariates Z that satisfies the backdoor criterion from X to Y in the model. Then, we regress Y on X and Z. The coefficient of X in the resulting equation represents the true causal effect of X on Y. The reasoning for this is similar to the reasoning we used to justify the backdoor criterion in the first place—regressing on Z adds those variables into the equation, blocking all backdoor paths from X and Y, thus preventing the coefficient of X from absorbing the spurious information those paths contain.

For example, consider a linear model that complies with the graph in Figure 3.14. If we want to find the total causal effect of X on Y, we first determine, using the backdoor criterion, that we must adjust for T. So we regress Y on X and T, using the regression equation $y = r_X X +$
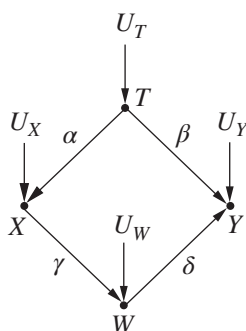

Figure 3.14   A graphical model in which X has no direct effect on Y, but a total effect that is determined by adjusting for T

$r_T T + \epsilon$. The coefficient r_X represents the total effect of X on Y. Note that this identification was possible without identifying any of the model parameters and without measuring variable W; the graph structure in itself gave us the license to ignore W, regress Y on T and X only, and identify the total effect (of X on Y) with the coefficient of X in that regression.

Suppose now that instead of the total causal effect, we want to find X's direct effect on Y. In a linear system, this direct effect is the structural coefficient α in the function $y = \alpha x + \beta z + \cdots + U_Y$ that defines Y in the system. We know from the graph of Figure 3.14 that $\alpha = 0$, because there is no direct arrow from X to Y. So, in this particular case, the answer is trivial: the direct effect is zero. But in general, how do we find the magnitude of α from data, if the model does not determine its value?
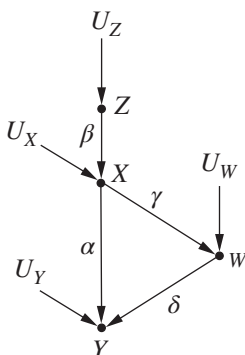

Figure 3.15   A graphical model in which X has direct effect α on Y

We can invoke a procedure similar to backdoor, except that now, we need to block not only backdoor paths but also indirect paths going from X to Y. First, we remove the edge from X to Y (if such an edge exists), and call the resulting graph G_α. If, in G_α, there is a set of variables Z that d-separates X and Y, then we can simply regress Y on X and Z. The coefficient of X in the resulting equation will equal the structural coefficient α.

The procedure above, which we might as well call "The Regression Rule for Identification" provides us with a quick way of determining whether any given parameter (say α) can be identified by ordinary least square (OLS) regression and, if so, what variables should go into the regression equation. For example, in the linear model of Figure 3.15, we can find the direct effect of X on Y by this method. First, we remove the edge between X and Y and get the graph G_α shown in Figure 3.16. It's easy to see that in this new graph, W d-separates X and Y. So we regress Y on X and W, using the regression equation $Y = r_X X + r_W W + \epsilon$. The coefficient r_X is the direct effect of X on Y.

Summarizing our observations thus far, two interesting features emerge. First, we see that, in linear systems, regression serves as the major tool for the identification and estimation of causal effects. To estimate a given effect, all we need to do is to write down a regression equation and specify (1) what variables should be included in the equation and (2) which of the coefficients in that equation represents the effect of interest. The rest is routine least square analysis on the sampled data which, as we remarked before, is facilitated by a variety of extremely efficient software packages. Second, we see that, as long as the U variables are independent of each
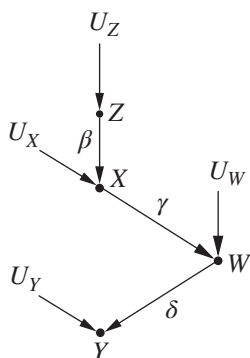
Figure 3.16 By removing the direct edge from X to Y and finding the set of variables $\{W\}$ that d-separate them, we find the variables we need to adjust for to determine the direct effect of X on Y

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
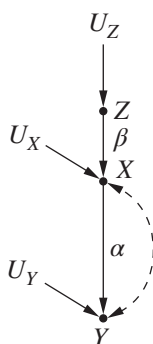

Figure 3.17 A graphical model in which we cannot find the direct effect of X on Y via adjustment, because the dashed double-arrow arc represents the presence of a backdoor path between X and Y, consisting of unmeasured variables. In this case, Z is an instrument with regard to the effect of X on Y that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

(a) *Identify three testable implications of this model.*

(b) *Identify a testable implication assuming that only X, Y, W_3, and Z_3 are observed.*

(c) *For each of the parameters in the model, write a regression equation in which one of the coefficients is equal to that parameter. Identify the parameters for which more than one such equation exists.*

(d) *Suppose X, Y, and W_3 are the only variables observed. Which parameters can be identified from the data? Can the total effect of X on Y be estimated?*

(e) *If we regress Z_1 on all other variables in the model, which regression coefficient will be zero?*

(f) *The model in Figure 3.18 implies that certain regression coefficients will remain invariant when an additional variable is added as a regressor. Identify five such coefficients with their added regressors.*

(g) *Assume that variables Z_2 and W_2 cannot be measured. Find a way to estimate b using regression coefficients. [Hint: Find a way to turn Z_1 into an instrumental variable for b.]*

3.8.4   Mediation in Linear Systems

When we can assume linear relationships between variables, mediation analysis becomes much simpler than the analysis conducted in nonlinear or nonparametric systems (Section 3.7). Estimating the direct effect of X on Y, for instance, amounts to estimating the path coefficient between the two variables, and this reduces to estimating correlation coefficients, using the techniques introduced in Section 3.8.3. The indirect effect, similarly, is computed via the difference $IE = \tau - DE$, where τ, the total effect, can be estimated by regression in the manner shown in Figure 3.14. In nonlinear systems, on the other hand, the direct effect is defined through expressions such as (3.18), or

$$DE = E[Y|do(x, z)] - E[Y|do(x', z)]$$

where $Z = z$ represents a specific stratum of all other parents of Y (besides X). Even when the identification conditions are satisfied, and we are able to reduce the $do()$ operators (by adjustments) to ordinary conditional expectations, the result will still depend on the specific values of x, x', and z. Moreover, the indirect effect cannot be given a definition in terms as do-expressions, since we cannot disable the capacity of Y to respond to X by holding variables constant. Nor can the indirect effect be defined as the difference between the total and direct effects, since differences do not faithfully reflect operations in nonlinear systems to X.

Such an operation will be introduced in Chapter 4 (Sections 4.4.5 and 4.5.2) using the language of counterfactuals.

Bibliographical Notes for chapter 3

Study question 3.3.2 is a version of Lord's paradox (Lord 1967), and is described in Glymour (2006), Hernández-Díaz et al. (2006), Senn (2006), and Wainer (1991). A unifying treatment is given in Pearl (2014b). The definition of the do-operator and "ACE" in terms of a modified model, has its conceptual origin with the economist Trygve Haavelmo (1943), who was the first

to simulate interventions by modifying equations in the model (see Pearl (2015c) for historical account). Strotz and Wold (1960) later advocated "wiping out" the equation determining *X*, and Spirtes et al. (1993) gave it a graphical representation in a form of a "manipulated graph." The "adjustment formula" of Eq. (3.5) as well as the "truncated product formula" first appeared in Spirtes et al. (1993), though these are implicit in the *G*-computation formula of Robins (1986), which was derived using counterfactual assumptions (see Chapter 4). The backdoor criterion of Definition 3.3.1 and its implications for adjustments were introduced in Pearl (1993). The front-door criterion and a general calculus for identifying causal effects from observations and experimental data (named *do*-calculus) were introduced in Pearl (1995) and further improved in Tian and Pearl (2002) and Shpitser and Pearl (2007), and Bareinboim and Pearl (2012). Section 3.7, and the identification of conditional interventions and *c*-specific effects is based on (Pearl 2009, pp. 113–114). Its extension to dynamic, time-varying policies is described in Pearl and Robins (1995) and (Pearl 2009, pp. 119–126). The role of covariate-specific effects in assessing interaction, moderation or effect modification is described in Morgan and Winship (2014) and Vanderweele (2015), whereas applications of Rule 2 to the detection of latent heterogeneity are described in Pearl (2015b). Additional discussions on the use of inverse probability weighting (Section 3.6) can be found in Hernán and Robins (2006). Our discussion of mediation (Section 3.7) and the identification of CDEs are based on Pearl (2009, pp. 126–130), whereas the fallibility of "conditioning" on a mediator to assess direct effects is demonstrated in Pearl (1998) as well as Cole and Hernán (2002).

The analysis of mediation has become extremely active in the past 15 years, primarily due to the advent of counterfactual logic (see Section 4.4.5); a comprehensive account of this progress is given in Vanderweele (2015). A tutorial survey of causal inference in linear systems (Section 3.8), focusing on parameter identification, is provided by Chen and Pearl (2014). Additional discussion on the confusion of regression versus structural equations can be found in Bollen and Pearl (2013).

A classic, and still the best textbook on the relationships between structural and regession coefficients is Heise (1975) (available online: http://www.indiana.edu/~socpsy/public_files/CausalAnalysis.zip). Other classics are Duncan (1975), Kenny (1979), and Bollen (1989). Classical texts, however, fall short of providing graphical tools of identification, such as those invoking backdoor and G_α (see Study question 3.8.1). A recent exception is Kline (2016).

Introductions to instrumental variables can be found in Greenland (2000) and in many textbooks of econometrics (e.g., Bowden and Turkington 1984, Wooldridge 2013). Generalized instrumental variables, extending the classical definition of Section 3.8.3 were introduced in Brito and Pearl (2002).

The program DAGitty (which is available online: http://www.dagitty.net/dags.html), permits users to search the graph for generalized instrumental variables, and reports the resulting IV estimators (Textor et al. 2011).

More recently, the *do*-calculus was used to solve problems of external validity, data-fusion, and meta-analysis (Bareinboim and Pearl 2013, Bareinboim and Pearl 2016, Pearl and Bareinboim 2014).

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

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

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, $P(driving\ time|do(freeway))$ and $P(driving\ time|do(Sepulveda))$, 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.

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

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

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 regimen, 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 $E_\wedge[Y_{X=1}|H \le H_0]$, 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., $H \le H_0$).

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 $E[Y|do(X = 1), Z = 1]$,
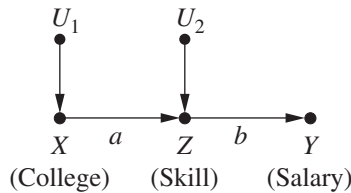

$U_1 \qquad U_2$

$X \quad a \quad Z \quad b \quad Y$
(College) (Skill) (Salary)

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

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

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

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

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

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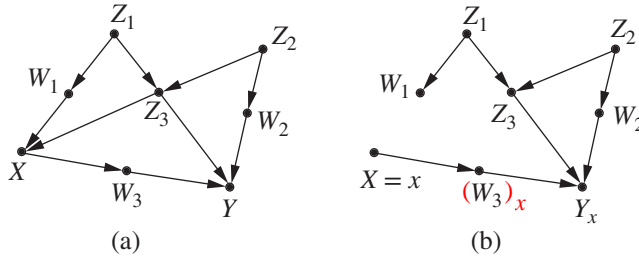
Figure 4.4   Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome Y predicated on $X = x$

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. 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 conditionalize on Z (as in Eq. (1.9)) and write

$$P(Y_x = y) = \sum_z P(Y_x = y | Z = z) P(z)$$

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

$$= \sum_z P(Y = y | Z = z, X = x) P(z) \tag{4.16}$$

(These variables remain the same in both models.)

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	Participant characteristics			Observed behavior			Predicted potential outcomes				
	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 $X = 1$

Participant	Predicted potential outcomes		Observed outcomes	
	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

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

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*

$$ETT = E[Y_1 - Y_0|X = 1] \qquad (4.18)$$

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

$$ETT = E[Y_1|X = 1] - E[Y_0|X = 1]$$
$$= E[Y|X = 1] - E[Y|X = 1] + \tau(1 - E[X|X = 1]) - \tau(0 - E[X|X = 1])$$
$$= \tau$$
$$= b + ac = 0.9$$

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 $ETT - \tau$ 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 of Model 2 (Figure 4.1) in 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

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

$$ETT = E[Y_1 - Y_0|X = 1]$$
$$= E[Y_1|X = 1] - E[Y_0|X = 1]$$
$$= E[Y|X = 1] - \sum_z E[Y|X = 0, Z = z]P(Z = z|X = 1)$$

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 $add(q)$:

$$E[Y|add(q)] - E[Y]$$

$$= \sum_{x'} E[Y_{x'+q}|X = x'] \frac{P(X = x')}{P(x')} - E[Y]$$

$$= \sum_{x'} \sum_z E[Y|X = x' + q, Z = z]P(Z = z|X = x')P(X = x') - E[Y] \qquad (4.22)$$

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

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

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:

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

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

$$PN = P(Y_0 = 0 | X = 1, Y = 1) \qquad (4.23)$$

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

$$PS = P(Y_1 = 1 | X = 0, Y = 0) \qquad (4.24)$$

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

$$PNS = \underset{0.39}{\overset{0.86 - 0.61}{}} - \underset{}{0.14} = 0.25$$

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 ~~sexuality~~ 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

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

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*

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

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

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

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

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

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

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

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

$$= \frac{0.002 - 0.028}{0.002} + \frac{0.028 - 0.014}{0.001} = -13 + 14 = 1 \qquad (4.41)$$

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

$$PNS = P(Y_x = 1, Y_{x'} = 0)$$

$$= P(Y_x = 1) - P(Y_{x'} = 1) \qquad (4.42)$$

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~~ 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:

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

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.

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

$$TE = NDE - NIE_r \qquad (4.48)$$

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 $T \to M$ and $T \to Y$).

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*

$$NDE = \sum_m \sum_w [E[Y|do(T = 1, M = m), W = w] - E[Y|do(T = 0, M = m), W = w]]$$

$$\times P(M = m|do(T = 0), W = w)P(W = w) \qquad (4.49)$$

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*

$$NDE = \sum_m \sum_w [E[Y|T = 1, M = m, W = w] - E[Y|T = 0, M = m, W = w]]$$

$$\times P(M = m|T = 0, W = w)P(W = w) \qquad (4.50)$$

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

$$NDE = \sum_m [E[Y|T = 1, M = m] - E[Y|T = 0, M = m]]P(M = m|T = 0). \qquad (4.51)$$

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

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

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 $E(Y \mid T = t, M = m)$
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 $E(M \mid T = t)$
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

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

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

$$TE = 0.80 \times 0.75 + 0.40 \times 0.25 - (0.30 \times 0.40 + 0.20 \times \overset{0.60}{\cancel{0.10}}) = 0.46$$

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

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.

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 Vander-Weele (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.

References

Balke A and Pearl J 1994a Counterfactual probabilities: Computational methods, bounds, and applications. In *Uncertainty in Artificial Intelligence 10* (ed. de Mantaras RL and Poole D) Morgan Kaufmann Publishers, San Mateo, CA pp. 46–54.

Balke A and Pearl J 1994b Probabilistic evaluation of counterfactual queries. *Proceedings of the Twelfth National Conference on Artificial Intelligence*, vol. **I**, MIT Press, Menlo Park, CA pp. 230–237.

Baron R and Kenny D 1986 The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology* **51** (6), 1173–1182.

Berkson J 1946 Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin* **2**, 47–53.

Bollen K 1989 *Structural Equations with Latent Variables*. John Wiley & Sons, Inc., New York.

Bollen K and Pearl J 2013 Eight myths about causality and structural equation models. In *Handbook of Causal Analysis for Social Research* (ed. Morgan S) Springer-Verlag, Dordrecht, Netherlands pp. 245–274.

Bowden R and Turkington D 1984 *Instrumental Variables*. Cambridge University Press, Cambridge, England.

Brito C and Pearl J 2002 Generalized instrumental variables. In *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference* (ed. Darwiche A and Friedman N) Morgan Kaufmann San Francisco, CA pp. 85–93.

Cai Z and Kuroki M 2006 Variance estimators for three 'probabilities of causation'. *Risk Analysis* **25** (6), 1611–1620.

Chen B and Pearl J 2014 *Graphical tools for linear structural equation modeling*. Technical Report R-432, Department of Computer Science, University of California, Los Angeles, CA. Submitted, Psychometrika, http://ftp.cs.ucla.edu/pub/stat_ser/r432.pdf.

Cole S and Hernán M 2002 Fallibility in estimating direct effects. *International Journal of Epidemiology* **31** (1), 163–165.

Conrady S and Jouffe L 2015 *Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers* 1st edition edn. Bayesia USA.

Cox D 1958 *The Planning of Experiments*. John Wiley and Sons, New York.

Darwiche A 2009 *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, New York.

Duncan O 1975 *Introduction to Structural Equation Models*. Academic Press, New York.

see page 131 for new refs added

Elwert F 2013 Graphical causal models. In *Handbook of Causal Analysis for Social Research* (ed. Morgan S) Springer-Verlag, Dordrecht, Netherlands pp. 245–274.

Fenton N and Neil M 2013 *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton, FL.

Fisher R 1922 On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 311.

Fisher B, Anderson S, Bryant J, Margolese RG, Deutsch M, Fisher ER, Jeong JH and Wolmark N 2002 Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *New England Journal of Medicine* **347** (16), 1233–1241.

Glymour MM 2006 Using causal diagrams to understand common problems in social epidemiology. *Methods in Social Epidemiology* John Wiley & Sons, Inc., San Francisco, CA pp. 393–428.

Glymour M and Greenland S 2008 Causal diagrams. In *Modern Epidemiology* (ed. Rothman K, Greenland S, and Lash T) 3rd edn. Lippincott Williams & Wilkins Philadelphia, PA pp. 183–209.

Greenland S 1999 Relation of probability of causation, relative risk, and doubling dose: A methodologic error that has become a social problem. *American Journal of Public Health* **89** (8), 1166–1169.

Greenland S 2000 An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology* **29** (4), 722–729.

Grinstead CM and Snell JL 1998 *Introduction to Probability* second revised edn. American Mathematical Society, United States.

Haavelmo T 1943 The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12. Reprinted in DF Hendry and MS Morgan (Eds.), 1995 *The Foundations of Econometric Analysis*, Cambridge University Press pp. 477–490.

Hayduk L, Cummings G, Stratkotter R, Nimmo M, Grygoryev K, Dosman D, Gilespie, M., Pazderka-Robinson H and Boadu K 2003 Pearl's *d*-separation: One more step into causal thinking. *Structural Equation Modeling* **10** (2), 289–311.

Heise D 1975 *Causal Analysis*. John Wiley and Sons, New York.

Hernán M and Robins J 2006 Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* **60** (7), 578–586, doi:10.1136/jech.2004.029496.

Hernández-Díaz S, Schisterman E and Hernán M 2006 The birth weight "paradox" uncovered? *American Journal of Epidemiology* **164** (11), 1115–1120.

Holland P 1986 Statistics and causal inference. *Journal of the American Statistical Association* **81** (396), 945–960.

Howard R and Matheson J 1981 Influence diagrams. In *Principles and Applications of Decision Analysis* (ed. Howard R and Matheson J) Strategic Decisions Group, Menlo Park, CA pp. 721-762.

Imai K, Keele L and Yamamoto T 2010 Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science* **25** (1), 51–71.

Jewell NP 2004 *Statistics for Epidemiology*. Chapman & Hall/CRC, Boca Raton, FL.

Kenny D 1979 *Correlation and Causality*. John Wiley & Sons, Inc., New York.

Kiiveri H, Speed T and Carlin J 1984 Recursive causal models. *Journal of Australian Math Society* **36**, 30–52.

Kim J and Pearl J 1983 A computational model for combined causal and diagnostic reasoning in inference systems. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, pp. 190–193, Karlsruhe, Germany.

Kline RB 2016 *Principles and Practice of Structural Equation Modeling* fourth: Revised and expanded edn. Guilford Publications, Inc., New York.

Koller K and Friedman N 2009 *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, United States.

Kyono T 2010 *Commentator: A front-end user-interface module for graphical and structural equation modeling*. Master's thesis Department of Computer Science, University of California, Los Angeles Los Angeles, CA.

Lauritzen S 1996 *Graphical Models*. Clarendon Press, Oxford. Reprinted 2004 with corrections.

Lewis D 1973 Causation. *Journal of Philosophy* **70**, 556–567.

Lindley DV 2014 *Understanding Uncertainty* revised edn. John Wiley & Sons, Inc., Hoboken, NJ.

Lord FM 1967 A paradox in the interpretation of group comparisons. *Psychological Bulletin* **68**, 304–305.

Mohan K, Pearl J and Tian J 2013 Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26* (ed. Burges C, Bottou L, Welling M, Ghahramani Z and Weinberger K) Neural Information Processing Systems Foundation, Inc. pp. 1277–1285.

Moore D, McCabe G and Craig B 2014 *Introduction to the Practice of Statistics*. W.H. Freeman & Co., New York.

Morgan SL and Winship C 2014 *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, *Analytical Methods for Social Research* 2nd edn. Cambridge University Press, New York.

Muthén B and Asparouhov T 2015 Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling: A Multidisciplinary Journal* **22** (1), 12-23.

Neyman J 1923 On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* **5** (4), 465–480.

Pearl J 1985 Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings, Cognitive Science Society*, pp. 329–334, Irvine, CA.

Pearl J 1986 Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* **29**, 241–288.

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

Pearl J 1993 Comment: Graphical models, causality, and intervention. *Statistical Science* **8** (3), 266–269.

Pearl J 1995 Causal diagrams for empirical research. *Biometrika* **82** (4), 669–710.

Pearl J 1998 Graphs, causality, and structural equation models. *Sociological Methods and Research* **27** (2), 226–284.

Pearl J 2000 *Causality: Models, Reasoning, and Inference* 2nd edn. Cambridge University Press, New York, 2009.

Pearl J 2001 Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* Morgan Kaufmann San Francisco, CA pp. 411–420.

Pearl J 2009 *Causality: Models, Reasoning, and Inference* 2nd edn. Cambridge University Press, New York.

Pearl J 2014a Interpretation and identification of causal mediation. *Psychological Methods* **19**, 459–481.

Pearl J 2016 *Lord's paradox revisited—(oh Lord! Kumbaya!)*. Technical Report R-436, Department of Computer Science, University of California, Los Angeles, CA. http://ftp.cs.ucla.edu/pub/stat_ser/r436.pdf. *Journal of Causal Inference* **4** (2) DOI: 10.1515/jci-2016-0021.

Pearl J 2014b Understanding Simpson's paradox. *The American Statistician* **88** (1), 8–13.

Pearl J 2015a Causes of effects and effects of causes. *Journal of Sociological Methods and Research* **44**, 149–164.

Pearl J 2015b Detecting latent heterogeneity. *Sociological Methods and Research* DOI: 10.1177/0049124115600597, online:1–20.

Pearl J and Bareinboim E 2014 External validity: From *do*-calculus to transportability across populations. *Statistical Science* **29**, 579–595.

Pearl J and Paz A 1987 GRAPHOIDS: A graph-based logic for reasoning about relevance relations In *Advances in Artificial Intelligence-II* (ed. Duboulay B, Hogg D and Steels L) North-Holland Publishing Co. pp. 357–363.

Pearl J and Robins J 1995 Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Uncertainty in Artificial Intelligence 11* (ed. Besnard P and Hanks S) Morgan Kaufmann, San Francisco, CA pp. 444–453.

Pearl J and Verma T 1991 A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference* (ed. Allena J, Fikes R and Sandewall E) Morgan Kaufmann San Mateo, CA pp. 441–452.

Pearl J 2015c Trygve Haavelmo and the emergence of causal calculus. *Econometric Theory*, Special issue on Haavelmo Centennial **31** (1), 152-179.

Pearl J and Mackenzie D 2018 *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York.

Pigou A 1911 *Alcoholism and Heredity*. Westminster Gazette. February 2.

Rebane G and Pearl J 1987 The recovery of causal poly-trees from statistical data. *Proceedings of the Third Workshop on Uncertainty in AI*, pp. 222–228, Seattle, WA.

Reichenbach H 1956 *The Direction of Time*. University of California Press, Berkeley, CA.

Robertson D 1997 The common sense of cause in fact. *Texas Law Review* **75** (7), 1765–1800.

Robins J 1986 A new approach to causal inference in mortality studies with a sustained exposure period—applications to control of the healthy workers survivor effect. *Mathematical Modeling* **7**, 1393–1512.

Robins J and Greenland S 1992 Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** (2), 143–155.

Rubin D 1974 Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Selvin S 2004 *Biostatistics: How it Works*. Pearson, New Jersey.

Senn S 2006 Change from baseline and analysis of covariance revisited. *Statistics in Medicine* **25**, 4334–4344.

Shpitser I 2013 Counterfactual graphical models for longitudinal mediation analysis with unobserved confounding. *Cognitive Science* **37** (6), 1011–1035.

Shpitser I and Pearl J 2007 What counterfactuals can be tested. *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence* AUAI Press Vancouver, BC, Canada pp. 352–359. Also, *Journal of Machine Learning Research* **9**, 1941–1979, 2008.

Shpitser I and Pearl J 2008 Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research* **9**, 1941–1979.

Shpitser I and Pearl J 2009 Effects of treatment on the treated: Identification and generalization. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* AUAI Press Montreal, Quebec pp. 514–521.

Simon H 1953 Causal ordering and identifiability. In *Studies in Econometric Method* (ed. Hood WC and Koopmans T) John Wiley & Sons, Inc. New York pp. 49–74.

Simpson E 1951 The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* **13**, 238–241.

Spirtes P and Glymour C 1991 An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review* **9** (1), 62–72.

Spirtes P, Glymour C and Scheines R 1993 *Causation, Prediction, and Search*. Springer-Verlag, New York.

Stigler SM 1999 *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, Hoboken, NJ.

Strotz R and Wold H 1960 Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica* **28**, 417–427.

Textor J, Hardt J and Knüüppel S 2011 DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology* **22** (5), 745.

Tian J, Paz A and Pearl J 1998 *Finding minimal d-separators*. Technical Report R-254, Department of Computer Science, University of California, Los Angeles, CA. http://ftp.cs.ucla.edu/pub/stat_ser/r254 .pdf.

Tian J and Pearl J 2000 Probabilities of causation: bounds and identification. *Annals of Mathematics and Artificial Intelligence* **28**, 287–313.

Tian J and Pearl J 2002 A general identification condition for causal effects. *Proceedings of the Eighteenth National Conference on Artificial Intelligence* AAAI Press/The MIT Press Menlo Park, CA pp. 567–573.

VanderWeele T 2015 *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.

Verma T and Pearl J 1988 Causal networks: Semantics and expressiveness. *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pp. 352–359, Mountain View, CA. Also in R.

Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4*, Elsevier Science Publishers, 69–76, 1990.

Verma T and Pearl J 1990 Equivalence and synthesis of causal models. *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 220–227, Cambridge, MA.

Virgil 29 BC Georgics. Verse 490, Book 2.

Wainer H 1991 Adjusting for differential base rates: Lord's paradox again. *Psychological Bulletin* **109**, 147–151.

Wooldridge J 2013 Introductory Econometrics: A Modern Approach 5th international edn. South-Western, Mason, OH.

NEW references - add to page 127:

Bareinboim E and Pearl J 2012 Causal inference by surrogate experiments (or, *z*-identifiability). *Proceedings of the Twenty-eighth Conference on Uncertainty in Artificial Intelligence* (ed. de Freitas N and Murphy K) AUAI Press, Corvallis, OR, pp. 113-120.

Bareinboim E and Pearl J 2013 A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* **1** (1), 107-134.

Bareinboim E and Pearl J 2016 Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* **113** (17), 7345-7352.

Bareinboim E, Tian J and Pearl J 2014 Recovering from selection bias in causal and statistical inference. *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* (ed. Brodley CE and Stone P) AAAI Press, Palo Alto, CA, pp. 2410-2416.