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: 5.17.17

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

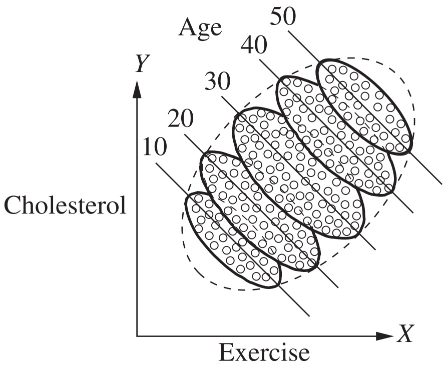


Figure 1.1 Results of the exercise–cholesterol study, segregated by age

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

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

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).

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

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

(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

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

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

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

$$y = r_0 + r_1x + r_2z \tag{1.23}$$

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,

$$Y = r_0 + r_1X_1 + r_2X_2 + \cdots + r_kX_k + \epsilon \tag{1.24}$$

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

→
$$X = \alpha + \beta Y + \epsilon \tag{1.25a}$$

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

→
$$E[X] = \alpha + \beta E[Y] \tag{1.25b}$$

→ Further multiplying both sides of (1.25a) by Y and taking the expectation gives

→
$$E[XY] = \alpha E[Y] + \beta E[Y^2] + E[Y\epsilon] \tag{1.26}$$

→ The orthogonality principle dictates $E[Y\epsilon] = 0$, and (1.25) and (1.26b) yield t wo equations with t wo unknowns, α and β. Solving f or 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.

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.

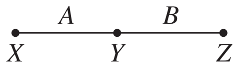


Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

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.

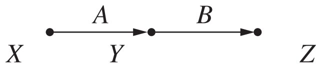


Figure 1.6 A directed graph in which node X is a parent of Y and Y is a parent of Z

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.

(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 2014c). 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 2.1.1 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 ~~—and, more importantly, that~~ 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

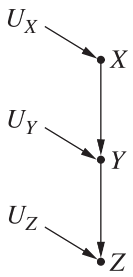


Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

its value; there is some case in which changing the value of the first variable changes the value of the second. That makes it likely 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~~ a typical 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)

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

$$f_X : X = U_X$$

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

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

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.

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 $4x, y, z, P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

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

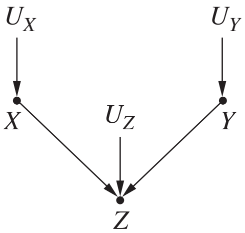


Figure 2.3 A simple collider

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 takes, for example, 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

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

X	Y	$P(X, Y \mid Z = 1)$
Heads	Heads	0.333
Heads	Tails	0.333
Tails	Heads	0.333
Tails	Tails	0
X	Y	$Pr(X, Y \mid Z = 0)$
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

$$P(X = \text{“}Heads\text{”} \mid Y = \text{“}Heads\text{”}) = P(X = \text{“}Heads\text{”}) = \frac{1}{2}$$

and

→

$$P(X = \text{“}Heads\text{”} \mid W = 1) = (0.25 + 0.25)\ \cancel{\text{or}}\ \div (0.25 + 0.25 + 0.25 + 0.125) = \frac{0.5}{0.85}$$

and

→

$$P(X = \text{“}Heads\text{”} \mid Y = \text{“}Heads\text{”},\ W = 1) = 0.25\ \cancel{\text{or}}\ \div (0.25 + 0.25) = 0.5 < \frac{0.5}{0.85}$$

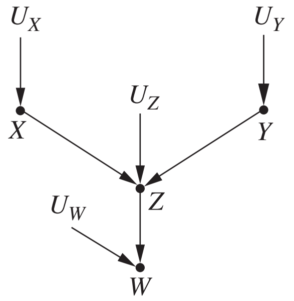


Figure 2.4 A simple collider, Z, with one child, W, representing the scenario from Table 2.3, with X representing one coin flip, Y representing the second coin flip, Z representing a bell that rings if either X or Y is heads, and W representing an unreliable witness who reports on whether or not the 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

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 2003; 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 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) and handling missing data (Mohan et al. 2013).

, recovering from sampling selection bias (Bareinboim et al. 2014), 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.)

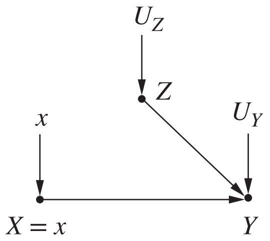


Figure 3.4 A modified graphical model representing an intervention on the model in Figure 3.3 that sets drug usage in the population, and results in the manipulated probability P_m

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

→

$$\begin{aligned} &P(Y = y|do(X = x)) \\ &\qquad = P_m(Y = y|X = x) \qquad \text{(by definition)} \end{aligned} \tag{3.2}$$

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

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

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*

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

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:

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

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 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.

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

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

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

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 and 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.

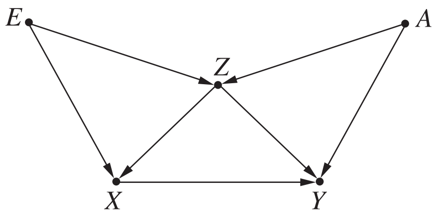


Figure 3.7 A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y

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 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

Bareinboim and Pearl 2012. Pearl 2009. Shpitser and Pearl 2008. and Tian and Pearl 2002 *effect is given by the following adjustment formula*

$$
\begin{aligned}
&P(Y=y|do(X=x),Z=z)\\
&\quad=\sum_{s}P(Y=y|X=x,S=s,Z=z)P(S=s\,|\,Z=z)
\end{aligned}
$$

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 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

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

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

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

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.

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
Yes	Yes	Female	0.547
Yes	No	Male	0.02
Yes	No	Female	0.202

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 $1/0.233$ 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:

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

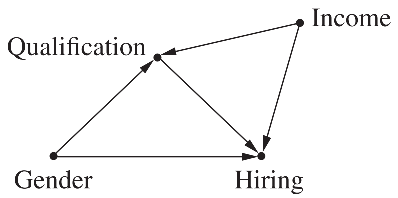
Table 3.5 Probability distribution for the population of Table 3.3 under the intervention $do(X = Yes)$, determined via the inverse probability method

X	Y	Z	% of population
Yes	Yes	Male	0.476
Yes	Yes	Female	0.357
Yes	No	Male	0.041
Yes	No	Female	0.132

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

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

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 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).

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 (named *do*-calculus) from observations and experimental data were introduced in Pearl (1995) and were 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 Pear1 (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.

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

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

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

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

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

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

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

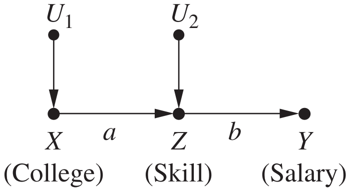


Figure 4.3 A model representing Eq. (4.7), illustrating the causal relations between college education (X), skills (Z), and salary (Y)

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

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

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

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

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

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

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

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

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

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)

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

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

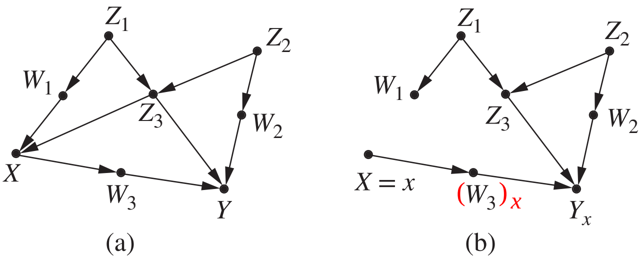


Figure 4.4 Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome Y predicated on $X = x$

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 covariate 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

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

(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

lower "|"s

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

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.

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 sample increases, the sample means converge to the population means.

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

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

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

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

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

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

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

Study questions

Study question 4.4.1

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

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

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

(c) *Repeat Question (b) using Theorem 4.3.2 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 reader may also wonder why $E[Y|add(q)]$ is not equal to the average causal effect

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

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?

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

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

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

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

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

Conditions for identifying natural effects

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

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

A-1 No member of W is a descendant of T.
A-2 W blocks all backdoor paths from M to Y (after removing $T \to M$ 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*

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

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

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

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

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

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

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

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

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

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

Numerical example: Mediation with binary variables

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

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

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

Treatment T	Homework M	Success rate $E(Y \mid T = t, M = m)$
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 ($Y = 1$) and untreated ($T = 0$) students

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

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.

see page 131 for new refs added

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. Forthcoming, 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.

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

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

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.

Replace ref

Muthén B ~~20~~ 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.
~~*SEM in M*~~
~~of California, Los Angeles, CA. Forthcoming, *Psychological Methods*.~~

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 2014b *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. Forthcoming, Journal of Causal Inference, 2016.

Pearl J 2014c 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.

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) In *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 In *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence* (ed. Brodley CE and Stone P) AAAI Press, Palo Alto, CA, pp. 2410-2416.