1

Preliminaries: Statistical and Causal Models

1.1 Why Study Causation

The answer to the question "why study causation?" is almost as immediate as the answer to "why study statistics." We study causation because we need to make sense of data, to guide actions and policies, and to learn from our success and failures. We need to estimate the effect of smoking on lung cancer, of education on salaries, of carbon emissions on the climate. Most ambitiously, we also need to understand *how* and *why* causes influence their effects, which is not less valuable. For example, knowing whether malaria is transmitted by mosquitoes or "mal-air," as many believed in the past, tells us whether we should pack mosquito nets or breathing masks on our next trip to the swamps.

Less obvious is the answer to the question, "why study causation as a separate topic, distinct from the traditional statistical curriculum?" What can the concept of "causation," considered on its own, tell us about the world that tried-and-true statistical methods can't?

Quite a lot, as it turns out. When approached rigorously, causation is not merely an aspect of statistics; it is an addition to statistics, an enrichment that allows statistics to uncover workings of the world that traditional methods alone cannot. For example, and this might come as a surprise to many, none of the problems mentioned above can be articulated in the standard language of statistics.

To understand the special role of causation in statistics, let's examine one of the most intriguing puzzles in the statistical literature, one that illustrates vividly why the traditional language of statistics must be enriched with new ingredients in order to cope with cause–effect relationships, such as the ones we mentioned above.

1.2 Simpson's Paradox

Named after Edward Simpson (born 1922), the statistician who first popularized it, the paradox refers to the existence of data in which a statistical association that holds for an entire population is reversed in every subpopulation. For instance, we might discover that students who

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

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

smoke get higher grades, on average, than nonsmokers get. But when we take into account the students' age, we might find that, in every age group, smokers get lower grades than nonsmokers get. Then, if we take into account both age and income, we might discover that smokers once again get *higher* grades than nonsmokers of the same age and income. The reversals may continue indefinitely, switching back and forth as we consider more and more attributes. In this context, we want to decide whether smoking causes grade increases and in which direction and by how much, yet it seems hopeless to obtain the answers from the data.

In the classical example used by Simpson (1951), a group of sick patients are given the option to try a new drug. Among those who took the drug, a lower percentage recovered than among those who did not. However, when we partition by gender, we see that *more* men taking the drug recover than do men are not taking the drug, and more women taking the drug recover than do women are not taking the drug! In other words, the drug appears to help men and women, but hurt the general population. It seems nonsensical, or even impossible—which is why, of course, it is considered a paradox. Some people find it hard to believe that numbers could even be combined in such a way. To make it believable, then, consider the following example:

Example 1.2.1 *We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in Table 1.1.*

The first row shows the outcome for male patients; the second row shows the outcome for female patients; and the third row shows the outcome for all patients, regardless of gender. In male patients, drug takers had a better recovery rate than those who went without the drug (93% vs 87%). In female patients, again, those who took the drug had a better recovery rate than nontakers (73% vs 69%). However, in the combined population, those who did not take the drug had a better recovery rate than those who did (83% vs 78%).

The data seem to say that if we know the patient's gender—male or female—we can prescribe the drug, but if the gender is unknown we should not! Obviously, that conclusion is ridiculous. If the drug helps men and women, it must help *anyone*; our lack of knowledge of the patient's gender cannot make the drug harmful.

Given the results of this study, then, should a doctor prescribe the drug for a woman? A man? A patient of unknown gender? Or consider a policy maker who is evaluating the drug's overall effectiveness on the population. Should he/she use the recovery rate for the general population? Or should he/she use the recovery rates for the gendered subpopulations?

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

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 gender (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

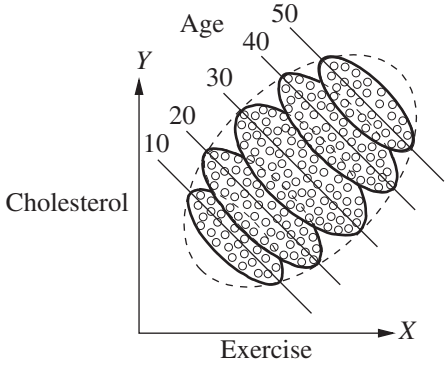

Figure 1.1 Results of the exercise–cholesterol study, segregated by age

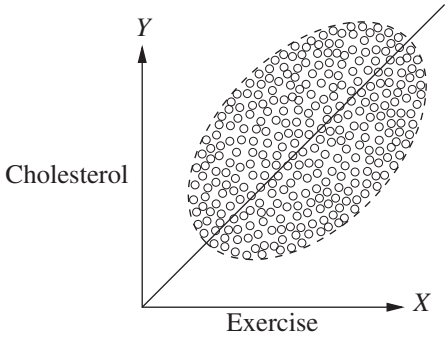

Figure 1.2 Results of the exercise–cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

recorded at the end of the experiment. In this case, we know that the drug affects recovery by lowering the blood pressure of those who take it—but unfortunately, it also has a toxic effect. At the end of our experiment, we receive the results shown in Table 1.2. (Table 1.2 is numerically identical to Table 1.1, with the exception of the column labels, which have been switched.)

Now, would you recommend the drug to a patient?

Once again, the answer follows from the way the data were generated. In the general population, the drug might improve recovery rates because of its effect on blood pressure. But in the subpopulations—the group of people whose posttreatment BP is high and the group whose posttreatment BP is low—we, of course, would not see that effect; we would only see the drug's toxic effect.

As in the gender example, the purpose of the experiment was to gauge the overall effect of treatment on rates of recovery. But in this example, since lowering blood pressure is one of the mechanisms by which treatment affects recovery, it makes no sense to separate the results based on blood pressure. (If we had recorded the patients' blood pressure *before* treatment, and if it were BP that had an effect on treatment, rather than the other way around, it would be a different story.) So we consult the results for the general population, we find that treatment increases the probability of recovery, and we decide that we *should* recommend treatment. Remarkably, though the numbers are the same in the gender and blood pressure examples, the correct result lies in the segregated data for the former and the aggregate data for the latter.

None of the information that allowed us to make a treatment decision—not the timing of the measurements, not the fact that treatment affects blood pressure, and not the fact that blood

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No drug	Drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

pressure affects recovery—was found in the data. In fact, as statistics textbooks have traditionally (and correctly) warned students, correlation is not causation, so there is no statistical method that can determine the causal story from the data alone. Consequently, there is no statistical method that can aid in our decision.

Yet statisticians interpret data based on causal assumptions of this kind all the time. In fact, the very paradoxical nature of our initial, qualitative, gender example of Simpson's problem is derived from our strongly held conviction that treatment cannot affect sex. If it could, there would be no paradox, since the causal story behind the data could then easily assume the same structure as in our blood pressure example. Trivial though the assumption "treatment does not cause sex" may seem, there is no way to test it in the data, nor is there any way to represent it in the mathematics of standard statistics. There is, in fact, no way to represent *any* causal information in contingency tables (such as Tables 1.1 and 1.2), on which statistical inference is often based.

There are, however, *extra*-statistical methods that can be used to express and interpret causal assumptions. These methods and their implications are the focus of this book. With the help of these methods, readers will be able to mathematically describe causal scenarios of any complexity, and answer decision problems similar to those posed by Simpson's paradox as swiftly and comfortably as they can solve for X in an algebra problem. These methods will allow us to easily distinguish each of the above three examples and move toward the appropriate statistical analysis and interpretation. A calculus of causation composed of simple logical operations will clarify the intuitions we already have about the nonexistence of a drug that cures men and women but hurts the whole population and about the futility of comparing patients with equal blood pressure. This calculus will allow us to move beyond the toy problems of Simpson's paradox into intricate problems, where intuition can no longer guide the analysis. Simple mathematical tools will be able to answer practical questions of policy evaluation as well as scientific questions of how and why events occur.

But we're not quite ready to pull off such feats of derring-do just yet. In order to rigorously approach our understanding of the causal story behind data, we need four things:

1. A working definition of "causation."
2. A method by which to formally articulate causal assumptions—that is, to create causal models.
3. A method by which to link the structure of a causal model to features of data.
4. A method by which to draw conclusions from the combination of causal assumptions embedded in a model and data.

The first two parts of this book are devoted to providing methods for modeling causal assumptions and linking them to data sets, so that in the third part, we can use those assumptions and data to answer causal questions. But before we can go on, we must define causation. It may seem intuitive or simple, but a commonly agreed-upon, completely encompassing definition of causation has eluded statisticians and philosophers for centuries. For our purposes, the definition of causation is simple, if a little metaphorical: A variable X is a *cause* of a variable Y if Y in any way relies on X for its value. We will expand slightly upon this definition later, but for now, think of causation as a form of listening; X is a cause of Y if Y listens to X and decides its value in response to what it hears.

Readers must also know some elementary concepts from probability, statistics, and graph theory in order to understand the aforementioned causal methods. The next two sections

will therefore provide the necessary definitions and examples. Readers with a basic understanding of probability, statistics, and graph theory may skip to Section 1.5 with no loss of understanding.

Study questions

Study question 1.2.1

What is wrong with the following claims?

(a) *"Data show that income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married."*
(b) *"Data show that as the number of fires increase, so does the number of fire fighters. Therefore, to cut down on fires, you should reduce the number of fire fighters."*
(c) *"Data show that people who hurry tend to be late to their meetings. Don't hurry, or you'll be late."*

Study question 1.2.2

A baseball batter Tim has a better batting average than his teammate Frank. However, someone notices that Frank has a better batting average than Tim against both right-handed and left-handed pitchers. How can this happen? (Present your answer in a table.)

Study question 1.2.3

Determine, for each of the following causal stories, whether you should use the aggregate or the segregated data to determine the true effect.

(a) *There are two treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore, more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data, or the stone size-specific data when determining which treatment will be more effective?*
(b) *There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery and the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately, to maximize the chance of a successful surgery?*

Study question 1.2.4

In an attempt to estimate the effectiveness of a new drug, a randomized experiment is conducted. In all, 50% of the patients are assigned to receive the new drug and 50% to receive a placebo. A day before the actual experiment, a nurse hands out lollipops to some patients who

show signs of depression, mostly among those who have been assigned to treatment the next day (i.e., the nurse's round happened to take her through the treatment-bound ward). Strangely, the experimental data revealed a Simpson's reversal: Although the drug proved beneficial to the population as a whole, drug takers were less likely to recover than nontakers, among both lollipop receivers and lollipop nonreceivers. Assuming that lollipop sucking in itself has no effect whatsoever on recovery, answer the following questions:

(a) Is the drug beneficial to the population as a whole or harmful?
(b) Does your answer contradict our gender example, where sex-specific data was deemed more appropriate?
(c) Draw a graph (informally) that more or less captures the story. (Look ahead to Section 1.4 if you wish.)
(d) How would you explain the emergence of Simpson's reversal in this story?
(e) Would your answer change if the lollipops were handed out (by the same criterion) a day after the study?

[Hint: Use the fact that receiving a lollipop indicates a greater likelihood of being assigned to drug treatment, as well as depression, which is a symptom of risk factors that lower the likelihood of recovery.]

1.3 Probability and Statistics

Since statistics generally concerns itself not with absolutes but with likelihoods, the language of probability is extremely important to it. Probability is similarly important to the study of causation because most causal statements are uncertain (e.g., "careless driving causes accidents," which is true, but does not mean that a careless driver is certain to get into an accident), and probability is the way we express uncertainty. In this book, we will use the language and laws of probability to express our beliefs and uncertainty about the world. To aid readers without a strong background in probability, we provide here a glossary of the most important terms and concepts they will need to know in order to understand the rest of the book.

1.3.1 Variables

A *variable* is any property or descriptor that can take multiple values. In a study that compares the health of smokers and nonsmokers, for instance, some variables might be the age of the participant, the gender of the participant, whether or not the participant has a family history of cancer, and how many years the participant has been smoking. A variable can be thought of as a question, to which the value is the answer. For instance, "How old is this participant?" "38 years old." Here, "age" is the variable, and "38" is its value. The probability that variable X takes value x is written $P(X = x)$. This is often shortened, when context allows, to $P(x)$. We can also discuss the probability of multiple values at once; for instance, the probability that $X = x$ and $Y = y$ is written $P(X = x, Y = y)$, or $P(x, y)$. Note that $P(X = 38)$ is specifically interpreted as the probability that an individual randomly selected from the population is aged 38.

A variable can be either *discrete* or *continuous*. Discrete variables (sometimes called *categorical* variables) can take one of a finite or countably infinite set of values in any range. A variable describing the state of a standard light switch is discrete, because it has two values: "on"

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

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

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

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

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

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

Study question 1.3.9

(a) Prove Eq. (1.22) using the orthogonality principle. [Hint: Follow the treatment of Eq. (1.26).]

(b) Find all partial regression coefficients

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

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

1.4 Graphs

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

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

object. A mathematical graph is a collection of *vertices* (or, as we will call them, *nodes*) and edges. The nodes in a graph are connected (or not) by the edges. Figure 1.5 illustrates a simple graph. X, Y, and Z (the dots) are nodes, and A and B (the lines) are edges.

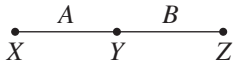

Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

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.

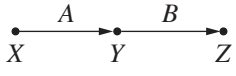

Figure 1.6 A directed graph in which node A is a parent of B and B is a parent of C

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) there is no directed path from any node to itself, whereas in (2) there are directed paths from X back to X, for example.

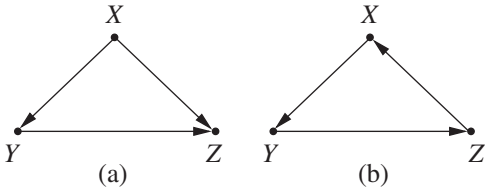

Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

Study questions

Study question 1.4.1

Consider the graph shown in Figure 1.8:

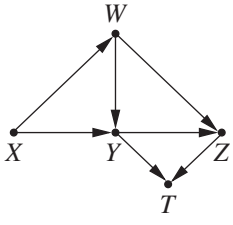

Figure 1.8 A directed graph used in Study question 1.4.1

(a) Name all of the parents of Z.
(b) Name all the ancestors of Z.
(c) Name all the children of W.
(d) Name all the descendants of W.
(e) Draw all (simple) paths between X and T (i.e., no node should appear more than once).
(f) Draw all the directed paths between X and T.

1.5 Structural Causal Models

1.5.1 Modeling Causal Assumptions

In order to deal rigorously with questions of causality, we must have a way of formally setting down our assumptions about the causal story behind a data set. To do so, we introduce the concept of the *structural causal model*, or SCM, which is a way of describing the relevant features of the world and how they interact with each other. Specifically, a structural causal model describes how nature assigns values to variables of interest.

Formally, a structural causal model consists of two sets of variables U and V, and a set of functions f that assigns each variable in V a value based on the values of the other variables in the model. Here, as promised, we expand on our definition of causation: A variable X is a *direct cause* of a variable Y if X appears in the function that assigns Y's value. X is a *cause* of Y if it is a direct cause of Y, or of any cause of Y.

The variables in U are called *exogenous variables*, meaning, roughly, that they are external to the model; we choose, for whatever reason, not to explain how they are caused. The variables in V are *endogenous*. Every endogenous variable in a model is a descendant of at least one exogenous variable. Exogenous variables cannot be descendants of any other variables, and in particular, cannot be a descendant of an endogenous variable; they have no ancestors and are represented as *root* nodes in graphs. If we know the value of every exogenous variable, then using the functions in f, we can determine with perfect certainty the value of every endogenous variable.

For example, suppose we are interested in studying the causal relationships between a treatment X and lung function Y for individuals who suffer from asthma. We might assume that Y also depends on, or is "caused by," air pollution levels as captured by a variable Z. In this case, we would refer to X and Y as endogenous and Z as exogenous. This is because we assume that air pollution is an external factor, that is, it cannot be caused by an individual's selected treatment or their lung function.

Every SCM is associated with a *graphical causal model*, referred to informally as a "graphical model" or simply "graph." Graphical models consist of a set of nodes representing the variables in U and V, and a set of edges between the nodes representing the functions in f. The graphical model G for an SCM M contains one node for each variable in M. If, in M, the function f_X for a variable X contains within it the variable Y (i.e., if X depends on Y for its value), then, in G, there will be a directed edge from Y to X. We will deal primarily with SCMs for which the graphical models are *directed acyclic graph*s (DAGs). Because of the relationship between SCMs and graphical models, we can give a graphical definition of causation: If, in a graphical model, a variable X is the child of another variable Y, then Y is a direct cause of X; if X is a descendant of Y, then Y is a potential cause of X (there are rare *intransitive cases* in which Y will not be a cause of X, which we will discuss in Part Two).

In this way, causal models and graphs encode causal assumptions. For instance, consider the following simple SCM:

SCM 1.5.1 (Salary Based on Education and Experience)

$$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$$
$$f_Z : Z = 2X + 3Y$$

This model represents the salary (Z) that an employer pays an individual with X years of schooling and Y years in the profession. X and Y both appear in f_Z, so X and Y are both direct causes of Z. If X and Y had any ancestors, those ancestors would be potential causes of Z.

The graphical model associated with SCM 1.5.1 is illustrated in Figure 1.9.

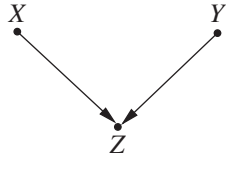

Figure 1.9 The graphical model of SCM 1.5.1, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary

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

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

SCM 1.5.2 (Basketball Performance Based on Height and Sex)

$$V = \{\text{Height, Sex, Performance}\}, \quad U = \{U_1, U_2, U_3\}, \quad F = \{f1, f2\}$$
$$\text{Sex} = U_1$$
$$\text{Height} = f_1(\text{Sex}, U_2)$$
$$\text{Performance} = f_2(\text{Height, Sex}, U_3)$$

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

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

Study questions

Study question 1.5.1

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

SCM 1.5.3

$$V = \{X, Y, Z\}, \quad U = \{U_X, U_Y, U_Z\}, \quad F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = u_X$$
$$f_Y : Y = \frac{X}{3} + U_Y$$
$$f_Z : Z = \frac{Y}{16} + U_Z$$

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

1.5.2 Product Decomposition

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

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

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

Rule of product decomposition

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

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

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

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

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

This knowledge allows us to save an enormous amount of space when laying out a joint distribution. We need not create a probability table that lists a value for every possible triple (x, y, z). It will suffice to create three much smaller tables for X, $(Y|X)$, and $(Z|Y)$, and multiply the values as necessary.

To estimate the joint distribution from a data set generated by the above model, we need not count the frequency of every triple; we can instead count the frequencies of each x, $(y|x)$, and $(z|y)$ and multiply. This saves us a great deal of processing time in large models. It also increases substantially the accuracy of frequency counting. Thus, the assumptions underlying the graph allow us to exchange a "high-dimensional" estimation problem for a few "low-dimensional" probability distribution challenges. The graph therefore simplifies an estimation problem and, simultaneously, provides more precise estimators. If we do not know the graphical structure of an SCM, estimation becomes impossible with large number of variables and small, or moderately sized, data sets—the so-called "curse of dimensionality."

Graphical models let us do all of this without always needing to know the functions relating the variables, their parameters, or the distributions of their error terms.

Here's an evocative, if unrigorous, demonstration of the time and space saved by this strategy: Consider the chain $X \to Y \to Z \to W$, where X stands for clouds/no clouds, Y stands for rain/no rain, Z stands for wet pavement/dry pavement, and W stands for slippery pavement/unslippery pavement.

Using your own judgment, based on your experience of the world, how plausible is it that $P(\textit{clouds, no-rain, dry pavement, slippery pavement}) = 0.23$?

This is quite a difficult question to answer straight out. But using the product rule, we can break it into pieces:

$$P(\textit{clouds})P(\textit{no rain}|\textit{clouds})P(\textit{dry pavement}|\textit{no rain})P(\textit{slippery pavement}|\textit{dry pavement})$$

Our general sense of the world tells us that $P(\textit{clouds})$ should be relatively high, perhaps 0.5 (lower, of course, for those of us living in the strange, weatherless city of Los Angeles). Similarly, $P(\textit{no rain}|\textit{clouds})$ is fairly high—say, 0.75. And $P(\textit{dry pavement}|\textit{no rain})$ would be higher still, perhaps 0.9. But the $P(\textit{slippery pavement}|\textit{dry pavement})$ should be quite low, somewhere in the range of 0.05. So putting it all together, we come to a ballpark estimate of $0.5 \times 0.75 \times 0.9 \times 0.05 = 0.0169$.

We will use this product rule often in this book in cases when we need to reason with numerical probabilities, but wish to avoid writing out large probability tables.

The importance of the product decomposition rule can be particularly appreciated when we deal with estimation. In fact, much of the role of statistics focuses on effective sampling designs, and estimation strategies, that allow us to exploit an appropriate data set to estimate probabilities as precisely as we might need. Consider again the problem of estimating the probability $P(X, Y, Z, W)$ for the chain $X \to Y \to Z \to W$. This time, however, we attempt to estimate the probability from data, rather than our own judgment. The number of (x, y, z, w) combinations that need to be assigned probabilities is $16 - 1 = 15$. Assume that we have 45 random observations, each consisting of a vector (x, y, z, w). On the average, each (x, y, z, w) cell would receive about three samples; some will receive one or two samples, and some remain empty. It is very unlikely that we would obtain a sufficient number of samples in each cell to assess the proportion in the population at large (i.e., when the sample size goes to infinity).

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

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

Study questions

Study question 1.5.2

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

(a) Based on this model, compute the joint distributions $P(x, y, z), P(x, y), P(x, z)$, and $P(y, z)$ for all values of x, y, and z, in terms of the parameters $(r, p_1, p_2, p_3, p_4, q_1, q_2)$. [Hint: Use the product decomposition of Section 1.5.2.]

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

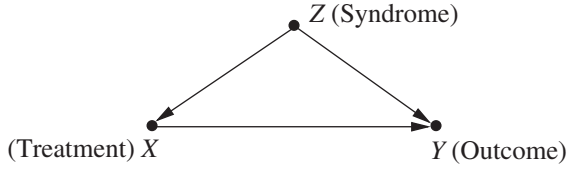

Figure 1.10 Model showing an unobserved syndrome, Z, affecting both treatment (X) and outcome (Y)

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

Study question 1.5.3

Consider a graph $X_1 \to X_2 \to X_3 \to X_4$ 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 paradoxis 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 ofstatistics, 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

basis of *Bayesian Networks* (Pearl 1985)—directed acyclic graphs that represent probabilistic knowledge, not necessarily causal. For inference and applications of Bayesian networks, see Darwiche (2009) and Fenton and Neil (2013), and Conrady and Jouffe (2015). The validity of the product decomposition rule for structural causal models was shown in Pearl and Verma (1991).