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 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

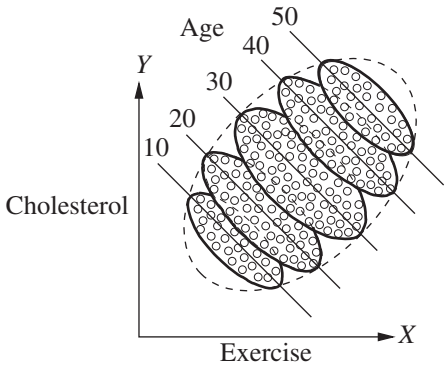

Figure 1.1 Results of the exercise–cholesterol study, segregated by age

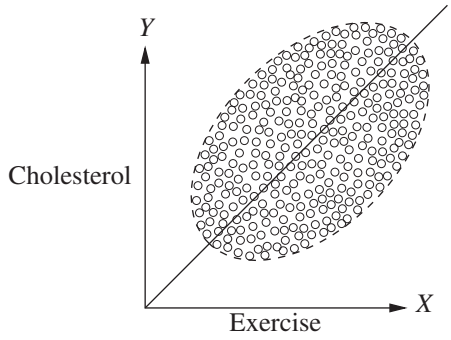

Figure 1.2 Results of the exercise–cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

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"

and "off." Continuous variables can take any one of an infinite set of values on a continuous scale (i.e., for any two values, there is some third value that lies between them). For instance, a variable describing in detail a person's weight is continuous, because weight is measured by a real number.

1.3.2 Events

An *event* is any assignment of a value or set of values to a variable or set of variables. "$X = 1$" is an event, as is "$X = 1$ or $X = 2$," as is "$X = 1$ and $Y = 3$," as is "$X = 1$ or $Y = 3$." "The coin flip lands on heads," "the subject is older than 40," and "the patient recovers" are all events. In the first, "outcome of the coin flip" is the variable, and "heads" is the value it takes. In the second, "age of the subject" is the variable and "older than 40" describes a set of values it may take. In the third, "the patient's status" is the variable and "recovery" is the value. This definition of "event" runs counter to our everyday notion, which requires that some change occur. (For instance, we would not, in everyday conversation, refer to a person being a certain age as an event, but we would refer to that person *turning* a year older as such.) Another way of thinking of an event in probability is this: Any declarative statement (a statement that can be true or false) is an event.

Study questions

Study question 1.3.1

Identify the variables and events invoked in the lollipop story of Study question 1.2.4

1.3.3 Conditional Probability

The probability that some event A occurs, given that we know some other event B has occurred, is the *conditional probability of A given B*. The conditional probability that $X = x$, given that $Y = y$, is written $P(X = x|Y = y)$. As with unconditional probabilities, this is often shortened to $P(x|y)$. Often, the probability that we assign to the event "$X = x$" changes drastically, depending on the knowledge "$Y = y$" that we condition on. For instance, the probability that you have the flu right now is fairly low. But, that probability would become much higher if you were to take your temperature and discover that it is 102 °F.

When dealing with probabilities represented by frequencies in a data set, one way to think of conditioning is *filtering* a data set based on the value of one or more variables. For instance, suppose we looked at the ages of U.S. voters in the last presidential election. According to the Census Bureau, we might get the data set shown in Table 1.3.

In Table 1.3, there were 132,948,000 votes cast in total, so we would estimate that the probability that a given voter was younger than the age of 45 is

$$P(\textit{Voter's Age} < 45) = \frac{20{,}539{,}000 + 30{,}756{,}000}{132{,}448{,}000} = \frac{51{,}295{,}000}{132{,}948{,}000} = 0.38$$

Suppose, however, we want to estimate the probability that a voter was younger than the age of 45, given that we *know* he was elder than the age of 29. To find this out, we simply filter

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).

Table 1.5 The proportion of males and females achieving a given education level

Gender	Highest education achieved	Occurrence (in hundreds of thousands)
Male	Never finished high school	112
Male	High school	231
Male	College	595
Male	Graduate school	242
Female	Never finished high school	136
Female	High school	189
Female	College	763
Female	Graduate school	172

1.3.4 Independence

It might happen that the probability of one event remains unaltered with the observation of another. For example, while observing your high temperature increases the probability that you have the flu, observing that your friend Joe is 38 years old does not change the probability at all. In cases such as this, we say that the two events are *independent*. Formally, events A and B are said to be independent if

$$P(A|B) = P(A) \tag{1.1}$$

that is, the knowledge that B has occurred gives us no additional information about the probability of A occurring. If this equality does not hold, then A and B are said to be *dependent*. Dependence and independence are symmetric relations—if A is dependent on B, then B is dependent on A, and if A is independent of B, then B is independent of A. (Formally, if $P(A|B) = P(A)$, then it *must* be the case that $P(B|A) = P(B)$.) This makes intuitive sense; if "smoke" tells us something about "fire," then "fire" must tell us something about "smoke."

Two events A and B are *conditionally independent* given a third event C if

$$P(A|B, C) = P(A|C) \tag{1.2}$$

and $P(B|A, C) = P(B|C)$. For example, the event "smoke detector is on" is dependent on the event "there is a fire nearby." But these two events may become independent conditional on the third event "there is smoke nearby"; smoke detectors respond to the presence of smoke only, not to its cause. When dealing with data sets, or probability tables, A and B are conditionally independent given C if A and B are independent in the new data set created by filtering on C. If A and B are independent in the original unfiltered data set, they are called *marginally independent*.

Variables, like events, can be dependent or independent of each other. Two variables X and Y are considered independent if for every value x and y that X and Y can take, we have

$$P(X = x|Y = y) = P(X = x) \tag{1.3}$$

(As with independence of events, independence of variables is a symmetrical relation, so it follows that Eq. (1.3) implies $P(Y = y|X = x) = P(Y = y)$.) If for any pair of values of X and Y,

this equality does not hold, then X and Y are said to be *dependent*. In this sense, independence of variables can be understood as a set of independencies of events. For instance, "height" and "musical talent" are independent variables; for every height h and level of musical talent m, the probability that a person is h feet high would not change upon discovering that he/she has m amount of talent.

1.3.5 Probability Distributions

A *probability distribution* for a variable X is the set of probabilities assigned to each possible value of X. For instance, if X can take three values—1, 2, and 3—a possible probability distribution for X would be "$P(X = 1) = 0.5, P(X = 2) = 0.25, P(X = 3) = 0.25$." The probabilities in a probability distribution must lie between 0 and 1, and must sum to 1. An event with probability 0 is impossible; an event with probability 1 is certain.

Continuous variables also have probability distributions. The probability distribution of a continuous variable X is represented by a function f, called the *density function*. When f is plotted on a coordinate plane, the probability that the value of variable X lies between values a and b is the area under the curve between a and b—or, as those who have taken calculus will know, $\int_a^b f(x)dx$. The area under the entire curve—that is, $\int_{-\infty}^{\infty} f(x)dx$—must of course be equal to 1.

Sets of variables can also have probability distributions, called *joint distributions*. The joint distribution of a set of variables V is the set of probabilities of each possible combination of variable values in V. For instance, if V is a set of two variables—X and Y—each of which can take two values—1 and 2—then one possible joint distribution for V is "$P(X = 1, Y = 1) = 0.2, P(X = 1, Y = 2) = 0.1, P(X = 2, Y = 1) = 0.5, P(X = 2, Y = 2) = 0.2$." Just as with single-variable distributions, probabilities in a joint distribution must sum to 1.

1.3.6 The Law of Total Probability

There are several universal probabilistic truths that are useful to know. First, for any two mutually exclusive events A and B (i.e., A and B cannot co-occur), we have

$$P(A \text{ or } B) = P(A) + P(B) \tag{1.4}$$

It follows that, for any two events A and B, we have

$$P(A) = P(A, B) + P(A, \textit{"not } B\textit{"}) \tag{1.5}$$

because the events "A and B" and "A and 'not B"' are mutually exclusive—and because if A is true, then either "A and B" or "A and 'not B"' must be true. For example, "Dana is a tall man" and "Dana is a tall woman" are mutually exclusive, and if Dana is tall, then he or she must be either a tall man or a tall woman; therefore, $P(\textit{Dana is tall}) = P(\textit{"Dana is a tall man"}) + P(\textit{"Dana is a tall woman"})$.

More generally, for any set of events $B_1, B_2, \ldots, B_n$ such that exactly one of the events must be true (an exhaustive, mutually exclusive set, called a *partition*), we have

$$P(A) = P(A, B_1) + P(A, B_2) + \cdots + P(A, B_n) \tag{1.6}$$

This rule, known as the *law of total probability*, becomes somewhat obvious as soon as we put it in real-world terms: If we pull a random card from a standard deck, the probability that the card is a Jack will be equal to the probability that it's a Jack *and* a spade, plus the probability that it's a Jack *and* a heart, plus the probability that it's a Jack *and* a club, plus the probability that it's a Jack *and* a diamond. Calculating the probability of an event A by summing up its probabilities over all B_i is called *marginalizing over* B, and the resulting probability $P(A)$ is called the *marginal probability* of A.

If we know the probability of B and the probability of A conditional on B, we can deduce the probability of A and B by simple multiplication:

$$P(A, B) = P(A|B)P(B) \tag{1.7}$$

For instance, the probability that Joe is funny and smart is equal to the probability that a smart person is funny, multiplied by the probability that Joe is smart. The division rule

$$P(A|B) = P(A, B)/P(B)$$

which is formally regarded as a definition of conditional probabilities, is justified by viewing conditioning as a filtering operation, as we have done in Tables 1.3 and 1.4. When we condition on B, we remove from the table all events that conflict with B. The resulting subtable, like the original, represents a probability distribution, and like all probability distributions, it must sum to one. Since the probabilities of the subtables rows in the original distribution summed to $P(B)$ (by definition), we can determine their probabilities in the new distribution by multiplying each by $1/P(B)$.

Equation (1.7) implies that the notion of independence, which until now we have used informally to mean "giving no additional information," has a numerical representation in the probability distribution. In particular, for events A and B to be independent, we require that

$$P(A, B) = P(A)P(B)$$

For example, to check if the outcomes of two coins are truly independent, we should count the frequency at which both show up tails, and make sure that it equals the product of the frequencies at which each of the coins shows up tails.

Using (1.7) together with the symmetry $P(A, B) = P(B, A)$, we can immediately obtain one of the most important laws of probability, *Bayes' rule*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1.8}$$

With the help of the multiplication rule in (1.7), we can express the law of total probability as a weighted sum of conditional probabilities:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k) \tag{1.9}$$

This is very useful, because often we will find ourselves in a situation where we cannot assess $P(A)$ directly, but we can through this decomposition. It is generally easier to assess conditional probabilities such as $P(A|B_k)$, which are tied to specific contexts, rather than $P(A)$, which is not attached to a context. For instance, suppose we have a stock of gadgets from two sources: 30% of them are manufactured by factory A, in which one out of 5000 is defective, whereas 70%

are manufactured by factory B, in which one out of 10,000 is defective. To find the probability that a randomly chosen gadget will be defective is not a trivial mental task, but when broken down according to Eq. (1.9) it becomes easy:

$$\begin{aligned} P(\textit{defective}) &= P(\textit{defective}|A)P(A) + P(\textit{defective}|B)P(B) \\ &= \frac{0.30}{5{,}000} + \frac{0.70}{10{,}000} \\ &= \frac{1.30}{10{,}000} = 0.00013 \end{aligned}$$

Or, to take a somewhat harder example, suppose we roll two dice, and we want to know the probability that the second roll is higher than the first, $P(A) = P(\textit{Roll}\ 2 > \textit{Roll}\ 1)$. There is no obvious way to calculate this probability all at once. But if we break it down into contexts $B_1, \ldots, B_6$ by conditioning on the value of the first die, it becomes easy to solve:

$$\begin{aligned} P(\textit{Roll}\ 2 > \textit{Roll}\ 1) &= P(\textit{Roll}\ 2 > \textit{Roll}\ 1|\textit{Roll}\ 1 = 1)P(\textit{Roll}\ 1 = 1) \\ &\quad + P(\textit{Roll}\ 2 > \textit{Roll}\ 1|\textit{Roll}\ 1 = 2)P(\textit{Roll}\ 1 = 2) \\ &\quad + \cdots + P(\textit{Roll}\ 2 > \textit{Roll}\ 1|\textit{Roll}\ 1 = 6) \times P(\textit{Roll}\ 1 = 6) \\ &= \left(\frac{5}{6} \times \frac{1}{6}\right) + \left(\frac{4}{6} \times \frac{1}{6}\right) + \left(\frac{3}{6} \times \frac{1}{6}\right) + \left(\frac{2}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{0}{6} \times \frac{1}{6}\right) \\ &= \frac{5}{12} \end{aligned}$$

The decomposition described in Eq. (1.9) is sometimes called "the law of alternatives" or "extending the conversation"; in this book, we will refer to it as *conditionalizing on B*.

1.3.7 Using Bayes' Rule

When using Bayes' rule, we sometimes loosely refer to event A as the "hypothesis" and event B as the "evidence." This naming reflects the reason that Bayes' theorem is so important: In many cases, we know or can easily determine $P(B|A)$ (the probability that a piece of evidence will occur, given that our hypothesis is correct), but it's much harder to figure out $P(A|B)$ (the probability of the hypothesis being correct, given that we obtain a piece of evidence). Yet the latter is the question that we most often want to answer in the real world; generally, we want to update our belief in some hypothesis, $P(A)$, after some evidence B has occurred, to $P(A|B)$. To precisely use Bayes' rule in this manner, we must treat each hypothesis as an event and assign to all hypotheses for a given situation a probability distribution, called a *prior*.

For example, suppose you are in a casino, and you hear a dealer shout "11!" You happen to know that the only two games played at the casino that would occasion that event are craps and roulette and that there are exactly as many craps games as roulette games going on at any moment. What is the probability that the dealer is working at a game of craps, given that he shouted "11?"

In this case, "craps" is our hypothesis, and "11" is our evidence. It's difficult to figure out this probability off-hand. But the reverse—the probability that an 11 will result in a given round of craps—is easy to calculate; it is specified by the game. Craps is a game in which gamblers bet

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

$$\begin{aligned} P(\text{“11”}) &= P(\text{“11”}|\textit{“craps”})P(\textit{“craps”}) + P(\text{“11”}|\textit{“roulette”})P(\textit{“roulette”}) \\ &= \frac{1}{2} \times \frac{1}{18} + \frac{1}{2} \times \frac{1}{38} = \frac{7}{171} \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

Study questions

Study question 1.3.3

Consider the casino problem described in Section 1.3.6

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

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

Study question 1.3.4

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

(a) Use your intuition to argue that the probability that the face-down side of the card is also black is $\frac{1}{2}$. Why might it be greater than $\frac{1}{2}$?
(b) Express the probabilities and conditional probabilities that you find easy to estimate (for example, $P(C_D = Black)$), in terms of the following variables:

$$I = \text{Identity of the card selected (Card 1, Card 2, or Card 3)}$$
$$C_D = \text{Color of the face-down side (Black, White)}$$
$$C_U = \text{Color of the face-up side (Black, White)}$$

Find the probability that the face-down side of the selected card is black, using your estimates above.
(c) Use Bayes' theorem to find the correct probability of a randomly selected card's back being black if you observe that its front is black?

Study question 1.3.5 (Monty Hall)

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

1.3.8 Expected Values

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

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

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

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

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

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

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

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

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

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

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

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

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

(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

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

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

1.3.9 Variance and Covariance

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

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

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

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

$$\begin{aligned} Var(X) &= ((23.5 - 31.5)^2 \times 0.41) + ((37 - 31.5)^2 \times 0.59) \\ &= (64 \times 0.41) + (30.25 \times .59) \\ &= 26.24 + 17.85 = 43.09 \text{ years}^2 \end{aligned}$$

while the standard deviation is

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

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

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

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

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

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

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

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

Study questions

Study question 1.3.6

(a) Prove that, in general, both σ_{XY} and ρ_{XY} vanish when X and Y are independent. [*Hint: Use Eqs. (1.16) and (1.17).*]
(b) Give an example of two variables that are highly dependent and, yet, their correlation coefficient vanishes.

Study question 1.3.7

Two fair coins are flipped simultaneously to determine the payoffs of two players in the town's casino. Player 1 wins a dollar if and only if at least one coin lands on head. Player 2 receives a dollar if and only if the two coins land on the same face. Let X stand for the payoff of Player 1 and Y for the payoff of Player 2.

(a) Find and describe the probability distributions

$$P(x), P(y), P(x, y), P(y|x) \text{ and } P(x|y)$$

(b) Using the descriptions in (a), compute the following measures:

$$E[X], E[Y], E[Y|X = x], E[X|Y = y]$$

$$Var(X), Var(Y), Cov(X, Y), \rho_{XY}$$

(c) Given that Player 2 won a dollar, what is your best guess of Player 1's payoff?
(d) Given that Player 1 won a dollar, what is your best guess of Player 2's payoff?
(e) Are there two events, $X = x$ and $Y = y$, that are mutually independent?

Study question 1.3.8

Compute the following theoretical measures of the outcome of a single game of craps (one roll of two independent dice), where X stands for the outcome of Die 1, Z for the outcome of Die 2, and Y for their sum.

(a)

$$E[X], E[Y], E[Y|X = x], E[X|Y = y], \textit{ for each value of } x \textit{ and } y, \textit{ and}$$

$$Var(X), Var(Y), Cov(X, Y), \rho_{XY}, Cov(X, Z)$$

Table 1.6 describes the outcomes of 12 craps games.
(b) Find the sample estimates of the measures computed in (a), based on the data from Table 1.6. [*Hint: Many software packages are available for doing this computation for you.*]
(c) Use the results in (a) to determine the best estimate of the sum, Y, given that we measured $X = 3$.

Table 1.6 Results of 12 rolls of two fair dice

	X Die 1	Z Die 2	Y Sum
Roll 1	6	3	9
Roll 2	3	4	7
Roll 3	4	6	10
Roll 4	6	2	8
Roll 5	6	4	10
Roll 6	5	3	8
Roll 7	1	5	6
Roll 8	3	5	8
Roll 9	6	5	11
Roll 10	3	5	8
Roll 11	5	3	8
Roll 12	4	5	9

(d) What is the best estimate of X, given that we measured $Y = 4$?
(e) What is the best estimate of X, given that we measured $Y = 4$ and $Z = 1$? Explain why it is not the same as in (d).

1.3.10 Regression

Often, in statistics, we wish to predict the value of one variable, Y, based on the value of another variable, X. For example, we may want to predict a student's height based on his age. We noted earlier that the best prediction of Y based on X is given by the conditional expectation $E[Y|X = x]$, at least in terms of mean-squared error. But this assumes that we know the conditional expectation, or can compute it, from the joint distribution $P(y, x)$. With *regression*, we make our prediction directly from the data. We try to find a formula, usually a linear function, that takes observed values of X as input and gives values of Y as output, such that the square error between the predicted and actual values of Y is minimized, on average.

We start with a scatter plot that takes every case in our data set and charts them on a coordinate plane, as shown in Figure 1.2. Our predictor, or input, variable goes on the x-axis, and the variable whose value we are predicting goes on the y-axis.

The least squares regression line is the line for which the sum of the squared vertical distances of the points on the scatter plot from the line is minimized. That is, if there are n data points (x, y) on our scatter plot, and for any data point (x_i, y_i), the value y'_i represents the value of the line $y = \alpha + \beta x$ at x_i, then the least squares regression line is the one that minimizes the value

$$\sum_i (y_i - y'_i)^2 = \sum_i (y_i - \alpha - \beta x_i)^2 \tag{1.18}$$

To see how the slope β relates to the probability distribution $P(x, y)$, suppose we play 12 successive rounds of craps, and get the results shown in Table 1.6. If we wanted to predict the sum Y of the die rolls based on the value of $X = Die\ 1$ alone, using the data in Table 1.6, we would use the scatter plot shown in Figure 1.3. For our craps example, the least squares

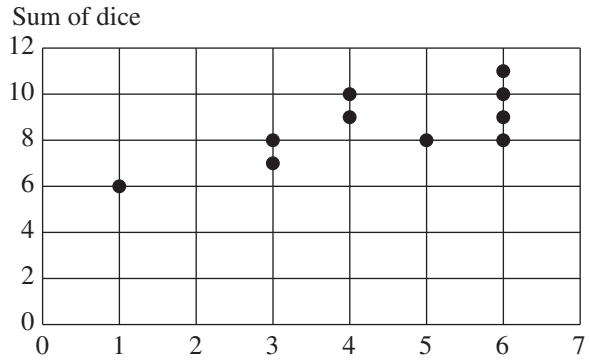

Figure 1.3 Scatter plot of the results in Table 1.6, with the value of Die 1 on the *x*-axis and the sum of the two dice rolls on the *y*-axis

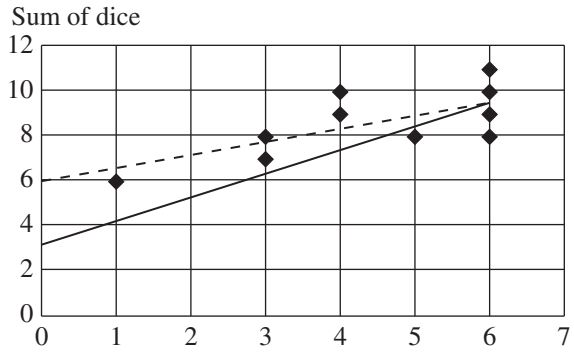

Figure 1.4 Scatter plot of the results in Table 1.6, with the value of Die 1 on the *x*-axis and the sum of the two dice rolls on the *y*-axis. The dotted line represents the line of best fit based on the data. The solid line represents the line of best fit we would expect in the population

regression line is shown in Figure 1.4. Note that the regression line for the *sample* that we used is not necessarily the same as the regression line for the *population*. The population is what we get when we allow our sample size to increase to infinity. The solid line in Figure 1.4 represents the theoretical least-square line, which is given by

$$y = 3.5 + 1.0x \tag{1.19}$$

The dashed line represents the sample least-square line, which, due to sampling variations, differs from the theoretical both in slope and in intercept.

In Figure 1.4, we know the equation of the regression line for the population because we know the expected value of the sum of two dice rolls, given that the first die lands on x. The computation is simple:

$$E[Y|X = x] = E[Die\ 2 + X|X = x] = E[Die\ 2] + x = 3.5 + 1.0x$$

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

$$Y = X + Z$$

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

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

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

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

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

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

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

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

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

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

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

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

1.3.11 Multiple Regression

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

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

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

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

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

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

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \tag{1.24}$$

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 (1.25b) and (1.26) yield two equations with two unknowns, α and β. Solving for α and β, we obtain

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

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

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

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

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

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

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

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

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

Study question 1.3.9

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

(b) Find all partial regression coefficients

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

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

1.4 Graphs

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

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

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

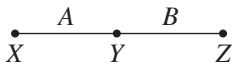

Figure 1.5 An undirected graph in which nodes X and Y are adjacent and nodes Y and Z are adjacent but not X and Z

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.

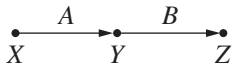

Figure 1.6 A directed graph in which node X is a parent of Y and Y is a parent of Z

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.

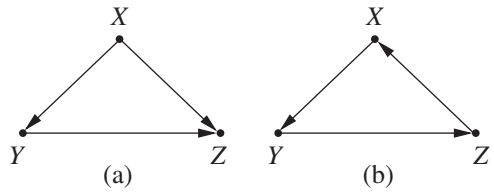

Figure 1.7 (a) Showing acyclic graph and (b) cyclic graph

Study questions

Study question 1.4.1

Consider the graph shown in Figure 1.8:

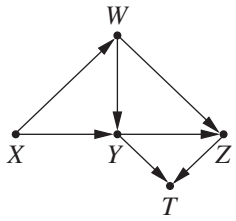

Figure 1.8 A directed graph used in Study question 1.4.1

(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.

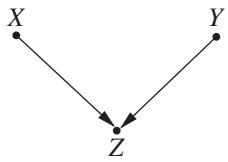

Figure 1.9 The graphical model of SCM 1.5.1, with X indicating years of schooling, Y indicating years of employment, and Z indicating salary

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* $E[Y|X = x, Z = z] = R_{YX \cdot Z}x + R_{YZ \cdot X}z$.]

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.

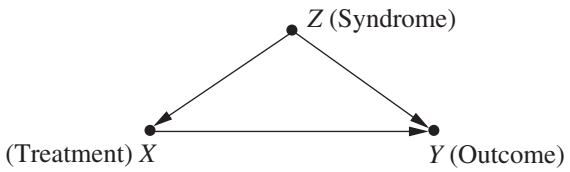

Figure 1.10 Model showing an unobserved syndrome, Z, affecting both treatment (X) and outcome (Y)

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

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

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).