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 it, 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 smoke get higher grades, on average, than nonsmokers. But when we take into account the students’ age, we might find that, in every age group, smokers get lower grades than nonsmokers. 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 recover than among those who did not. However, when we partition by gender, we see that more men taking the drug recover than do men not taking the drug, and more women taking the drug recover than do women not taking the drug! In other words, the drug appears to help men and help 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. 350 patients chose to take the drug and 350 patients did not. The results of the study are in Table 1.1.

<table>
<thead>
<tr>
<th>Drug</th>
<th>Men</th>
<th>Women</th>
<th>Combined data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>81 out of 87 recovered (93%)</td>
<td>192 out of 263 recovered (73%)</td>
<td>273 out of 350 recovered (78%)</td>
</tr>
<tr>
<td>No drug</td>
<td>234 out of 270 recovered (87%)</td>
<td>55 out of 80 recovered (69%)</td>
<td>289 out of 350 recovered (83%)</td>
</tr>
</tbody>
</table>

black 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 (93% vs. 87%). In female patients, again, those who took the drug had a better recovery rate than non-takers (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 should prescribe the drug, but if the gender is unknown we should not! Obviously, that conclusion is ridiculous. If the drug helps men and helps 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?

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. Additionally, 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 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 goes 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 (see 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.

![Figure 1.1: Results of the exercise-cholesterol study, segregated by age](image)

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, but instead of recording participants’ gender, patients’ blood pressure were 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
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.

1.2. (Table 1.2 is numerically identical to Table 1.1, with the exception of the column labels, which have been switched.)

<table>
<thead>
<tr>
<th></th>
<th>No drug</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low BP</td>
<td>81 out of 87 recovered (93%)</td>
<td>234 out of 270 recovered (87%)</td>
</tr>
<tr>
<td>High BP</td>
<td>192 out of 263 recovered (73%)</td>
<td>55 out of 80 recovered (69%)</td>
</tr>
<tr>
<td>Combined data</td>
<td>273 out of 350 recovered (78%)</td>
<td>289 out of 350 recovered (83%)</td>
</tr>
</tbody>
</table>

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 post-treatment 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 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 of treatment on recovery.

(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 deciding which treatment they would like to request?

(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 in choosing which surgeon to perform your operation.

**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