CHAPTER 1

Thinking Clearly in a Data-Driven Age

What You'll Learn

- Learning to think clearly and conceptually about quantitative information is important for lots of reasons, even if you have no interest in a career as a data analyst.
- Even well-trained people often make crucial errors with data.
- Thinking and data are complements, not substitutes.
- The skills you learn in this book will help you use evidence to make better decisions in your personal and professional life and be a more thoughtful and well-informed citizen.

Introduction

We live in a data-driven age. According to former Google CEO Eric Schmidt, the contemporary world creates as much new data every two days as had been created from the beginning of time through the year 2003. All this information is supposed to have the power to improve our lives, but to harness this power we must learn to think clearly about our data-driven world. Clear thinking is hard—especially when mixed up with all the technical details that typically surround data and data analysis.

Thinking clearly in a data-driven age is, first and foremost, about staying focused on ideas and questions. Technicality, though important, should serve those ideas and questions. Unfortunately, the statistics and quantitative reasoning classes in which most people learn about data do exactly the opposite—that is, they focus on technical details. Students learn mathematical formulas, memorize the names of statistical procedures, and start crunching numbers without ever having been asked to think clearly and conceptually about what they are doing or why they are doing it. Such an approach can work for people to whom thinking mathematically comes naturally. But we believe it is counterproductive for the vast majority of us. When technicality pushes students to stop thinking and start memorizing, they miss the forest for the trees. And it's also no fun.

Our focus, by contrast, is on conceptual understanding. What features of the world are you comparing when you analyze data? What questions can different kinds of comparisons answer? Do you have the right question and comparison for the problem you are trying to solve? Why might an answer that sounds convincing actually

be misleading? How might you use creative approaches to provide a more informative answer?

It isn't that we don't think the technical details are important. Rather, we believe that technique without conceptual understanding or clear thinking is a recipe for disaster. In our view, once you can think clearly about quantitative analysis, and once you understand why asking careful and precise questions is so important, technique will follow naturally. Moreover, this way is more fun.

In this spirit, we've written this book to require no prior exposure to data analysis, statistics, or quantitative methods. Because we believe conceptual thinking is more important, we've minimized (though certainly not eliminated) technical material in favor of plain-English explanations wherever possible. Our hope is that this book will be used as an introduction and a guide to how to think about and do quantitative analysis. We believe anyone can become a sophisticated consumer (and even producer) of quantitative information. It just takes some patience, perseverance, hard work, and a firm resolve to never allow technicality to be a substitute for clear thinking.

Most people don't become professional quantitative analysts. But whether you do or do not, we are confident you will use the skills you learn in this book in a variety of ways. Many of you will have quantitative analysts working for or with you. And all of you will read studies, news reports, and briefings in which someone tries to convince you of a conclusion using quantitative analyses. This book will equip you with the clear thinking skills necessary to ask the right questions, be skeptical when appropriate, and distinguish between useful and misleading evidence.

Cautionary Tales

To whet your appetite for the hard work ahead, let's start with a few cautionary tales that highlight the importance of thinking clearly in a data-driven age.

Abe's Hasty Diagnosis

Ethan's first child, Abe, was born in July 2006. As a baby, he screamed and cried almost non-stop at night for five months. Abe was otherwise happy and healthy, though a bit on the small side. When he was one year old the family moved to Chicago, without which move, you'd not be reading this book. (That last sentence contains a special kind of claim called a *counterfactual.* Counterfactuals are really important, and you are going to learn all about them in chapter 3.) After noticing that Abe was small for his age and growing more slowly than expected, his pediatrician decided to run some tests.

After some lab work, the doctors were pretty sure Abe had celiac disease—a digestive disease characterized by gluten intolerance. The good news: celiac disease is not life threatening or even terribly serious if properly managed through diet. The bad news: in 2007, the gluten-free dietary options for kids were pretty miserable.

It turns out that Abe actually had two celiac-related blood tests. One came back positive (indicating that he had the disease), the other negative (indicating that he did not have the disease). According to the doctors, the positive test was over 80 percent accurate. "This is a strong diagnosis," they said. The suggested course of action was to put Abe on a gluten-free diet for a couple of months to see if his weight increased. If it did, they could either do a more definitive biopsy or simply keep Abe gluten-free for the rest of his life.

Ethan asked for a look at the report on Abe's bloodwork. The doctors indicated they didn't think that would be useful since Ethan isn't a doctor. This response was neither

surprising nor hard to understand. People, especially experts and authority figures, often don't like acknowledging the limits of their knowledge. But Ethan wanted to make the right decision for his son, so he pushed hard for the information. One of the goals of this book is to give you some of the skills and confidence to be your own advocate in this way when using information to make decisions in your life.

Two numbers characterize the effectiveness of any diagnostic test. The first is its false negative rate, which is how frequently the test says a sick person is healthy. The second is its false positive rate, which is how frequently the test says a healthy person is sick. You need to know *both* the false positive rate and the false negative rate to interpret a diagnostic test's results. So Abe's doctors' statement that the positive blood test was 80 percent accurate wasn't very informative. Did that mean it had a 20 percent false negative rate? A 20 percent false positive rate? Do 80 percent of people who test positive have celiac disease?

Fortunately, a quick Google search turned up both the false positive and false negative rates for both of Abe's tests. Here's what Ethan learned. The test on which Abe came up positive for celiac disease has a false negative rate of about 20 percent. That is, if 100 people with celiac disease took the test, about 80 of them would correctly test positive and the other 20 would incorrectly test negative. This fact, we assume, is where the claim of 80 percent accuracy came from. The test, however, has a false positive rate of 50 percent! People who don't have celiac disease are just as likely to test positive as they are to test negative. (This test, it is worth noting, is no longer recommended for diagnosing celiac disease.) In contrast, the test on which Abe came up negative for celiac disease had much lower false negative and false positive rates.

Before getting the test results, a reasonable estimate of the probability of Abe having celiac disease, given his small size, was around 1 in 100. That is, about 1 out of every 100 small kids has celiac disease. Armed with the lab reports and the false positive and false negative rates, Ethan was able to calculate how likely Abe was to have celiac disease given his small size and the test results. Amazingly, the combination of testing positive on an inaccurate test and testing negative on an accurate test actually meant that the evidence suggested that Abe was much *less* likely than 1 in 100 to have celiac disease. In fact, as we will show you in chapter 15, the best estimate of the likelihood of Abe having celiac, given the test results, was about 1 in 1,000. The blood tests that Abe's doctors were sure supported the celiac diagnosis actually strongly supported the opposite conclusion. Abe was almost certain not to have celiac disease.

Ethan called the doctors to explain what he'd learned and to suggest that moving his pasta-obsessed son to a gluten-free diet, perhaps for life, was not the prudent next step. Their response: "A diagnosis like this can be hard to hear." Ethan found a new pediatrician.

Here's the upshot. Abe did not have celiac disease. The kid was just a bit small. Today he is a normal-sized kid with a ravenous appetite. But if his father didn't know how to think about quantitative evidence or lacked the confidence to challenge a mistaken expert, he'd have spent his childhood eating rice cakes. Rice cakes are gross, so he might still be small.

Civil Resistance

As many around the world have experienced, citizens often find themselves in deep disagreement with their government. When things get bad enough, they sometimes decide to organize protests. If you ever find yourself doing such organizing, you will face many important decisions. Perhaps none is more important than whether to build

a movement with a non-violent strategy or one open to a strategy involving more violent forms of confrontation. In thinking through this quandry, you will surely want to consult your personal ethics. But you might also want to know what the evidence says about the costs and benefits of each approach. Which kind of organization is most likely to succeed in changing government behavior? Is one or the other approach more likely to land you in prison, the hospital, or the morgue?

There is some quantitative evidence that you might use to inform your decisions. First, comparing anti-government movements across the globe and over time, governments more often make concessions to fully non-violent groups than to groups that use violence. And even comparing across groups that do use violence, governments more frequently make concessions to those groups that engage in violence against military and government targets rather than against civilians. Second, the personal risks associated with violent protest are greater than those associated with non-violent protest. Governments repress violent uprisings more often than they do non-violent protests, making concerns about prison, the hospital, and the morgue more acute.

This evidence sounds quite convincing. A non-violent strategy seems the obvious choice. It is apparently both more effective and less risky. And, indeed, on the basis of this kind of data, political scientists Erica Chenoweth and Evan Perkoski conclude that "planning, training, and preparation to maintain nonviolent discipline is key—especially (and paradoxically) when confronting brutal regimes."

But let's reconsider the evidence. Start by asking yourself, In what kind of a setting is a group likely to engage in non-violent rather than violent protest? A few thoughts occur to us. Perhaps people are more likely to engage in non-violent protest when they face a government that they think is particularly likely to heed the demands of its citizens. Or perhaps people are more likely to engage in non-violent protest when they have broad-based support among their fellow citizens, represent a group in society that can attract media attention, or face a less brutal government.

If any of these things are true, we should worry about the claim that maintaining non-violent discipline is key to building a successful anti-government movement. (Which isn't to say that we are advocating violence.) Let's see why.

Empirical studies find that, on average, governments more frequently make concessions in places that had non-violent, rather than violent, protests. The claimed implication rests on a particular interpretation of that difference—namely, that the higher frequency of government concessions in non-violent places is *caused* by the use of non-violent tactics. Put differently, all else held equal, if a given movement using violent methods had switched to using non-violent methods, the government would have been more likely to grant concessions. But is this causal interpretation really justified by the evidence?

Suppose it's the case that protest movements are more likely to turn to violence when they do not have broad-based support among their fellow citizens. Then, when we compare places that had violent protests to places that had non-violent protests, all else (other than protest tactics) is not held equal. Those places differ in at least two ways. First, they differ in terms of whether they had violent or non-violent protests. Second, they differ in terms of how supportive the public was of the protest movement.

This second difference is a problem for the causal interpretation. You might imagine that public opinion has an independent effect on the government's willingness to grant concessions. That is, all else held equal (including protest tactics), governments might be more willing to grant concessions to protest movements with broad-based public support. If this is the case, then we can't really know whether the fact that governments

grant concessions more often to non-violent protest movements than to violent protest movements is because of the difference in protest tactics or because the non-violent movements also happen to be the movements with broad-based public support. This is the classic problem of mistaking correlation for causation.

It is worth noting a few things. First, if government concessions are in fact due to public opinion, then it could be the case that, were we actually able to hold all else equal in our comparison of violent and non-violent protests, we would find the opposite relationship—that is, that non-violence is not more effective than violence (it could even be less effective). Given this kind of evidence, we just can't know.

Second, in this example, the conclusion that appears to follow if you don't force yourself to think clearly is one we would all like to be true. Who among us would not like to live in a world where non-violence is always preferred to violence? But the whole point of using evidence to help us make decisions is to force us to confront the possibility that the world may not be as we believe or hope it is. Indeed, it is in precisely those situations where the evidence seems to say exactly what you would like it to say that it is particularly important to force yourself to think clearly.

Third, we've pointed to one challenge in assessing the effects of peaceful versus violent protest, but there are others. For instance, think about the other empirical claim we discussed: that violent protests are more likely to provoke the government into repressive crack-downs than are non-violent protests. Recall, we suggested that people might be more likely to engage in non-violent protest when they are less angry at their government, perhaps because the government is less brutal. Ask yourself why, if this is true, we have a similar problem of interpretation. Why might the fact that there are more government crack-downs following violent protests than non-violent protests *not* mean that switching from violence to non-violence will reduce the risk of crack-downs? The argument follows a similar logic to the one we just made regarding concessions. If you don't see how the argument works yet, that's okay. You will by the end of chapter 9.

Broken-Windows Policing

In 1982, the criminologist George L. Kelling and the sociologist James Q. Wilson published an article in *The Atlantic* proposing a new theory of crime and policing that had enormous and long-lasting effects on crime policy in the United States and beyond.

Kelling and Wilson's theory is called *broken windows*. It was inspired by a program in Newark, New Jersey, that got police out of their cars and walking a beat. According to Kelling and Wilson, the program reduced crime by elevating "the level of public order." Public order is important, they argue, because its absence sets in motion a vicious cycle:

> A piece of property is abandoned, weeds grow up, a window is smashed. Adults stop scolding rowdy children... Families move out, unattached adults move in. Teenagers gather in front of the corner store. The merchant asks them to move; they refuse. Fights occur. Litter accumulates. People start drinking in front of the grocery...
>
> Residents will think that crime, especially violent crime, is on the rise... They will use the streets less often... Such an area is vulnerable to criminal invasion.

This idea that policing focused on minimizing disorder can reduce violent crime had a big impact on police tactics. Most prominently, the broken-windows theory was the

guiding philosophy in New York City in the 1990s. In a 1998 speech, then New York mayor Rudy Giuliani said,

> We have made the "Broken Windows" theory an integral part of our law enforcement strategy...
>
> You concentrate on the little things, and send the clear message that this City cares about maintaining a sense of law and order...then the City as a whole will begin to become safer.

And, indeed, crime in New York city did decline when the police started focusing "on the little things." According to a study by Hope Corman and H. Naci Mocan, misdemeanor arrests increased 70 percent during the 1990s and violent crime decreased by more than 56 percent, double the national average.

To assess the extent to which broken-windows policing was responsible for this fall in crime, Kelling and William Sousa studied the relationship between violent crime and broken-windows approaches across New York City's precincts. If minimizing disorder causes a reduction in violent crime, they argued, then we should expect the largest reductions in crime to have occurred in neighborhoods where the police were most focused on the broken-windows approach. And this is just what they found. In precincts where misdemeanor arrests (the "little things") were higher, violent crime decreased more. They calculated that "the average NYPD precinct...could expect to suffer one less violent crime for approximately every 28 additional misdemeanor arrests."

This sounds pretty convincing. But let's not be too quick to conclude that arresting people for misdemeanors is the answer to ending violent crime. Two other scholars, Bernard Harcourt and Jens Ludwig, encourage us to think a little more clearly about what might be going on in the data.

The issue that Harcourt and Ludwig point out is something called *reversion to the mean* (which we'll talk about a lot more in chapter 8). Here's the basic concern. In any given year, the amount of crime in a precinct is determined by lots of factors, including policing, drugs, the economy, the weather, and so on. Many of those factors are unknown to us. Some of them are fleeting; they come and go across precincts from year to year. As such, in any given precinct, we can think of there being some "baseline" level of crime, with some years randomly having more crime and some years randomly having less (relative to that precinct-specific baseline).

In any given year, if a precinct had a high level of crime (relative to its baseline), then it had bad luck on the unknown and fleeting factors that help cause crime. Probably next year its luck won't be as bad (that's what *fleeting* means), so that precinct will likely have less crime. And if a precinct had a low level of crime (relative to its baseline) this year, then it had good luck on the unknown and fleeting factors, and it will probably have worse luck next year (crime will go back up). Thus, year to year, the crime level in a precinct tends to revert toward the *mean* (i.e., the precinct's baseline level of crime).

Now, imagine a precinct that had a really high level of violent crime in the late 1980s. Two things are likely to be true of that precinct. First, it is probably a precinct with a high baseline of violent crime. Second, it is also probably a precinct that had a bad year or two—that is, for idiosyncratic and fleeting reasons, the level of crime in the late 1980s was high relative to that precinct's baseline. The same, of course, is true in reverse for precincts that had a low level of crime in the late 1980s. They probably have a low baseline of crime, and they also probably had a particularly good couple of years.

Why is this a problem for Kelling and Sousa's conclusions? Because of reversion to the mean, we would expect the most violent precincts in the late 1980s to show a reduction in violent crime on average, even with no change in policing. And unsurprisingly, given the police's objectives, but unfortunately for the study, it was precisely those high-crime precincts in the 1980s that were most likely to get broken-windows policing in the early 1990s. So, when we see a reduction in violent crime in the precincts that had the most broken-windows policing, we don't know if it's the policing strategy or reversion to the mean that's at work.

Harcourt and Ludwig go a step further to try to find more compelling evidence. Roughly speaking, they look at how changes in misdemeanor arrests relate to changes in violent crime in precincts that had similar levels of violent crime in the late 1980s. By comparing precincts with similar starting levels of violent crime, they go some way toward eliminating the problem of reversion to the mean. Surprisingly, this simple change actually flips the relationship! Rather than confirming Kelling and Sousa's finding that misdemeanor arrests are associated with a reduction in violent crime, Harcourt and Ludwig find that precincts that focused more on misdemeanor arrests actually appear to have experienced an *increase* in violent crime. Exactly the opposite of what we would expect if the broken-windows theory is correct.

Now, this reversal doesn't settle the matter on the efficacy of broken-windows policing. The relationship between misdemeanor arrests and violent crime that Harcourt and Ludwig find could be there for lots of reasons other than misdemeanor arrests causing an increase in violent crime. For instance, perhaps the neighborhoods with increasing misdemeanors are becoming less safe in general and would have experienced more violent crime regardless of policing strategies. What these results do show is that the data, properly considered, certainly don't offer the kind of unequivocal confirmation of the broken-windows ideas that you might have thought from Kelling and Sousa's finding. And you can only see this if you have the ability to think clearly about some subtle issues.

This flawed thinking was important. Evidence-based arguments like Kelling and Sousa's played a role in convincing politicians and policy makers that broken-windows policing was the right path forward when, in fact, it might have diverted resources away from preventing and investigating violent crime and may have created a more adversarial and unjust relationship between the police and the disproportionately poor and minority populations who were frequently cited for "the small stuff."

Thinking and Data Are Complements, Not Substitutes

Our quantitative world is full of lots of exciting new data and analytic tools to analyze that data with fancy names like machine learning algorithms, artificial intelligence, random forests, and neural networks. Increasingly, we are even told that this new technology will make it possible for the machines to do the thinking for us. But that isn't right. As our cautionary tales highlight, no data analysis, no matter how futuristic its name, will work if we aren't asking the right questions, if we aren't making the right comparisons, if the underlying assumptions aren't sound, or if the data used aren't appropriate. Just because an argument contains seemingly sophisticated quantitative data analysis, that doesn't mean the argument is rigorous or right. To harness the power of data to make better decisions, we must combine quantitative analysis with clear thinking.

Our stories also illustrate how our intuitions can lead us astray. It takes lots of care and practice to train ourselves to think clearly about evidence. The doctors'

intuition that Abe had celiac disease because of a test with 80 percent accuracy and the researchers' intuition that broken-windows policing works because crime decreased in places where it was deployed seem sensible. But both intuitions were wrong, suggesting that we should be skeptical of our initial hunches. The good news is that clear thinking can become intuitive if you work at it.

Data and quantitative tools are not a substitute for clear thinking. In fact, quantitative skills without clear thinking are quite dangerous. We suspect, as you read the coming chapters, you will be jarred by the extent to which unclear thinking affects even the most important decisions people make. Through the course of this book, we will see how misinterpreted information distorts life-and-death medical choices, national and international counterterrorism policies, business and philanthropic decisions made by some of the world's wealthiest people, how we set priorities for our children's education, and a host of other issues from the banal to the profound. Essentially, no aspect of life is immune from critical mistakes in understanding and interpreting quantitative information.

In our experience, this is because unclear thinking about evidence is deeply ingrained in human psychology. Certainly our own intuitions, left unchecked, are frequently subject to basic errors. Our guess is that yours are too. Most disturbingly, the experts on whose advice you depend—be they doctors, business consultants, journalists, teachers, financial advisors, scientists, or what have you—are often just as prone to making such errors as the rest of us. All too often, because they are experts, we trust their judgment without question, and so do they. That is why it is so important to learn to think clearly about quantitative evidence for yourself. That is the only way to know how to ask the right questions that lead you, and those on whose advice you depend, to the most reliable and productive conclusions possible.

How could experts in so many fields make important errors so often? Expertise, in any area, comes from training, practice, and experience. No one expects to become an expert in engineering, finance, plumbing, or medicine without instruction and years of work. But, despite its fundamental and increasing importance for so much of life in our quantitative age, almost no one invests this kind of effort into learning to think clearly with data. And, as we've said, even when they do, they tend to be taught in a way that over-emphasizes the technical and under-emphasizes the conceptual, even though the fundamental problems are almost always about conceptual mistakes in thinking rather than technical mistakes in calculation.

The lack of expertise in thinking presents us with two challenges. First, if so much expert advice and analysis is unreliable, how do you know what to believe? Second, how can you identify those expert opinions that do in fact reflect clear thinking?

This book provides a framework for addressing these challenges. Each of the coming chapters explains and illustrates, through a variety of examples, fundamental principles of clear thinking in a data-driven world. Part 1 establishes some shared language—clarifying what we mean by correlation and causation and what each is useful for. Part 2 discusses how we can tell whether a statistical relationship is genuine. Part 3 discusses how we can tell if that relationship reflects a causal phenomenon or not. And part 4 discusses how we should and shouldn't incorporate quantitative information into our decision making.

Our hope is that reading this book will help you internalize the principles of clear thinking in a deep enough way that they start to become second nature. You will know you are on the right path when you find yourself noticing basic mistakes in how people think and talk about the meaning of evidence everywhere you turn—as you watch

the news, peruse magazines, talk to business associates, visit the doctor, listen to the color commentary during athletic competitions, read scientific studies, or participate in school, church, or other communal activities. You will, we suspect, find it difficult to believe how much nonsense you're regularly told by all kinds of experts. When this starts to happen, try to remain humble and constructive in your criticisms. But do feel free to share your copy of this book with those whose arguments you find are in particular need of it. Or better yet, encourage them to buy their own copy!

Readings and References

The essay on non-violent protest by Erica Chenoweth and Evan Perkoski that we quote can be found at https://politicalviolenceataglance.org/2018/05/08/states-are-far-less-likely-to-engage-in-mass-violence-against-nonviolent-uprisings-than-violent-uprisings/.

The following book contains more research on the relationship between non-violence and efficacy:

Erica Chenoweth and Maria J. Stephan. 2011. *Why Civil Resistance Works: The Strategic Logic of Nonviolent Conflict.* Columbia University Press.

The following articles were discussed in this order on the topic of broken windows policing:

George L. Kelling and James Q. Wilson. 1982. "Broken Windows: The Police and Neighborhood Safety." *The Atlantic.* March https://www.theatlantic.com/magazine/archive/1982/03/broken-windows/304465/.

Archives of Rudolph W. Giuliani. 1998. "The Next Phase of Quality of Life: Creating a More Civil City." February 24. http://www.nyc.gov/html/rwg/html/98a/quality.html.

Hope Corman and H. Naci Mocan. 2005. "Carrots, Sticks, and Broken Windows." *Journal of Law and Economics* 48(1):235–66.

George L. Kelling and William H. Sousa, Jr. 2001. Do Police Matter? An Analysis of the Impact of New York City's Police Reforms. Civic Report for the Center for Civic Innovation at the Manhattan Institute.

Bernard E. Harcourt and Jens Ludwig. 2006. "Broken Windows: New Evidence from New York City and a Five-City Social Experiment." *University of Chicago Law Review* 73:271–320. *The published version has a misprinted sign in the key table. For the correction, see Errata, 74* U. Chi. L. Rev. 407 (2007).

PART I

Establishing a Common Language

CHAPTER 2

Correlation: What Is It and What Is It Good For?

What You'll Learn

- Correlations tell us about the extent to which two features of the world tend to occur together.
- In order to measure correlations, we must have data with variation in both features of the world.
- Correlations *can* be potentially useful for description, forecasting, and causal inference. But we have to think clearly about when they're appropriate for each of these tasks.
- Correlations are about linear relationships, but that's not as limiting as you might think.

Introduction

Correlation doesn't imply causation. That's a good adage. However, in our experience, it's less useful than it might be because, while many people know that correlation doesn't imply causation, hardly anyone knows what correlation and causation are.

In part 1, we are going to spend some time establishing a shared vocabulary. Making sure that we are all using these and a few other key terms to mean the same thing is absolutely critical if we are to think clearly about them in the chapters to come.

This chapter is about correlation: what it is and what it's good for. Correlation is the primary tool through which quantitative analysts describe the world, forecast future events, and answer scientific questions. Careful analysts do not avoid or disregard correlations. But they must think clearly about which kinds of questions correlations can and cannot answer in different situations.

What Is a Correlation?

The *correlation* between two features of the world is the extent to which they tend to occur together. This definition tells us that a correlation is a relationship between two things (which we call *features of the world* or *variables*). If two features of the world tend to occur together, they are *positively correlated.* If the occurrence of one feature of the world is unrelated to the occurrence of another feature of the world, they are *uncorrelated.* And if when one feature of the world occurs the other tends not to occur, they are *negatively correlated.*

Table 2.1. Oil production and type of government.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

What does it mean for two features of the world to tend to occur together? Let's start with an example of the simplest kind. Suppose we want to assess the correlation between two features of the world, and there are only two possible values for each one (we call these *binary* variables). For instance, whether it is after noon or before noon is a binary variable (by contrast, the time measured in hours, minutes, and seconds is not binary; it can take many more than two values).

Political scientists and economists sometimes talk about the *resource curse* or the *paradox of plenty*. The idea is that countries with an abundance of natural resources are often less economically developed and less democratic than those with fewer natural resources. Natural resources might make a country less likely to invest in other forms of development, or they might make a country more subject to violence and autocracy.

To assess the extent of this resource curse, we might want to know the correlation between natural resources and some feature of the economic or political system. That process starts with collecting some data, which we've done. To measure natural resources we looked at which countries are major oil producers. We classify a country as a major oil producer if it exports more than forty thousand barrels per day per million people. And for the political system we looked at which countries are considered autocracies versus democracies by the Polity IV Project. Table 2.1 indicates how many countries fit into each of the four possible categories: democracy and major oil producer, democracy and not major oil producer, autocracy and major oil producer, and autocracy and not major oil producer.

We can figure out if these two binary variables—being a major oil producer or not and autocracy versus democracy—are correlated by making a comparison. For instance, we could ask whether major oil producers are more likely to be autocracies than countries that aren't major oil producers. Or, similarly, we could ask whether autocracies are more likely to be major oil producers than democracies. If one of these statements is true, the other must be true as well. And these comparisons tell us whether these two features of the world—being a major oil producer and being an autocracy—tend to occur together.

In table 2.1, oil production and autocracy are indeed positively correlated. Fifty-five percent of major oil producers are autocracies ($\frac{11}{20} = .55$) while only about 20 percent of countries that aren't major oil producers are autocracies ($\frac{29}{147} \approx .20$). Equivalently, 27.5 percent of autocracies are major oil producers ($\frac{11}{40} = .275$), while only about 7 percent of democracies are ($\frac{9}{127} \approx .07$). In other words, major oil producers are more likely to be autocracies than are countries that aren't major oil producers, and then, necessarily, autocracies are more likely to be major oil producers than democracies.

As a descriptive matter, we find this positive correlation interesting. It is also potentially useful for prediction. Suppose there were some other countries outside our data

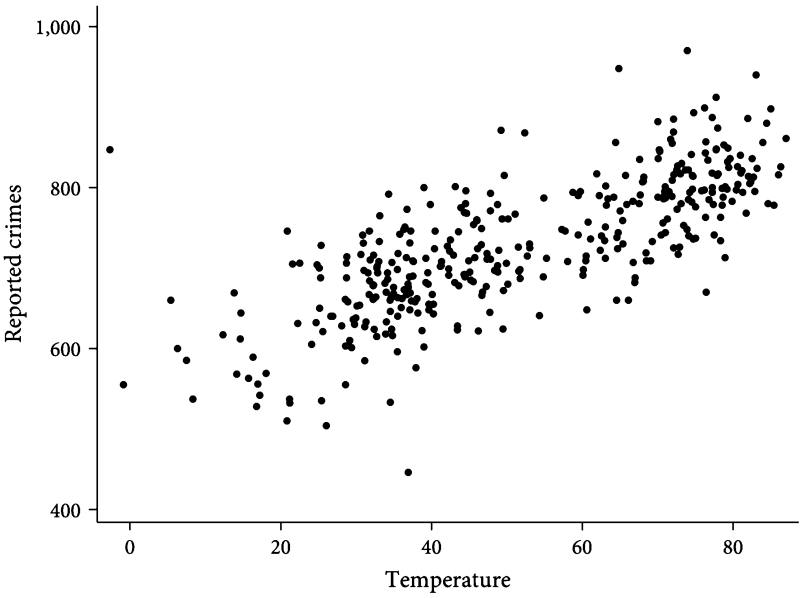

Figure 2.1. Crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

whose system of government we were uncertain of. Knowing whether or not they were major oil producers would be helpful in predicting which kind of government they likely have.

Such knowledge could even potentially be useful for causal inference. Perhaps new oil reserves are discovered in a country and the State Department wants to know what effect this is likely to have on the country's political system. This kind of data might be informative about that causal question as well. Though, as we'll discuss in great detail in chapter 9, we must be very careful when giving correlations this sort of causal interpretation.

We can assess correlations even when our data are such that it is hard to make a table of all the possible combinations like we did above. Suppose, for example, that we want to assess the relationship between crime and temperature in Chicago. We could assemble a spreadsheet in which each row corresponds to a day and each column corresponds to a feature of each day. We often call the rows *observations* and the features listed in the columns *variables*. In this case, the observations are different days. One variable could be the average temperature on that day as measured at Midway Airport. Another could be the number of crimes reported in the entire city of Chicago on that day. Another still could indicate whether the *Chicago Tribune* ran a story about crime on its front page on that day. As you can see, variables can take values that are binary (front page story or not), discrete but not binary (number of crimes), or continuous (average temperature).

We collected data like this for Chicago in 2018, and we'd like to assess the correlation between crime and temperature. But how can we assess the correlation between two non-binary variables?

One starting point is to make a simple graph, called a *scatter plot*. Figure 2.1 shows one for our 2018 Chicago data. In it, each point corresponds to an observation in our data—here, that means each point is a day in Chicago in 2018. The horizontal axis of our figure is the average temperature at Midway Airport on that day. The vertical axis

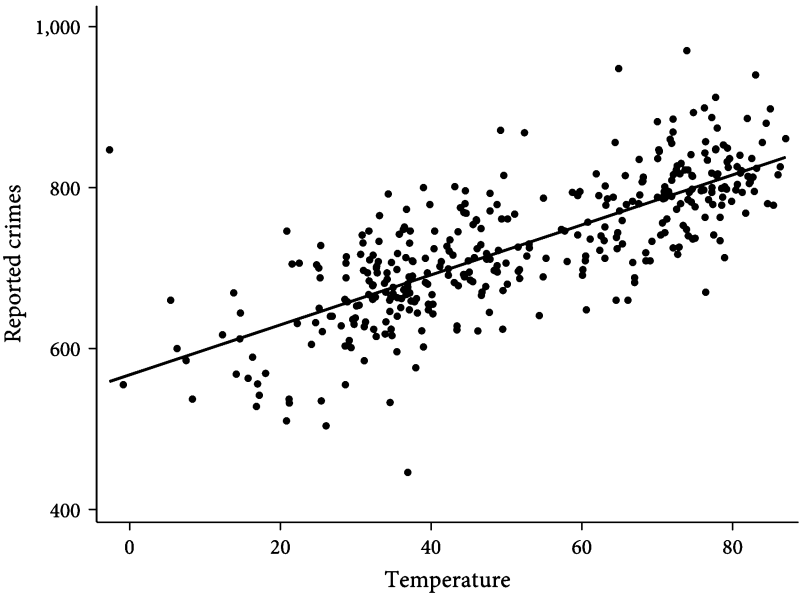

Figure 2.2. A line of best fit summarizing the relationship between crime and temperature (in degrees Fahrenheit) in Chicago across days in 2018.

is the number of crimes reported in the city on that day. So the location of each point shows the average temperature and the amount of crime on a given day.

Just by looking at the figure, you can see that it appears that there is a positive correlation between temperature and crime. Points to the left of the graph on the horizontal axis (colder days) tend to also be pretty low on the vertical axis (lower crime days), and days to right of the graph on the horizontal axis (warmer days) tend to also be pretty high on the vertical axis (higher crime days).

But how can we quantify this visual first impression? There are actually many different statistics that we can use to do so. One such statistic is called the *slope*. Suppose we found *the line of best fit* for the data. By *best fit*, we mean, roughly, the line that minimizes how far the data points are from the line on average. (We will be more precise about this in chapter 5.) The slope of the line of best fit is one way of describing the correlation between these two continuous variables.

Figure 2.2 shows the scatter plot with that line added. The slope of the line tells us something about the relationship between those two variables. If the slope is negative, the correlation is negative. If the slope is zero, temperature and crime are uncorrelated. If the slope is positive, the correlation is positive. And the steepness of the slope tells us about the strength of the correlation between these two variables. Here we see that they are positively correlated—there tends to be more crime on warmer days. In particular, the slope is 3.1, so on average for every additional degree of temperature (in Fahrenheit), there are 3.1 more crimes.

Notice that how you interpret the slope depends on which variable is on the vertical axis and which one is on the horizontal axis. Had we drawn the graph the other way around (as in figure 2.3), we would be describing the relationship between the same two variables. But this time, we would have learned that for every additional

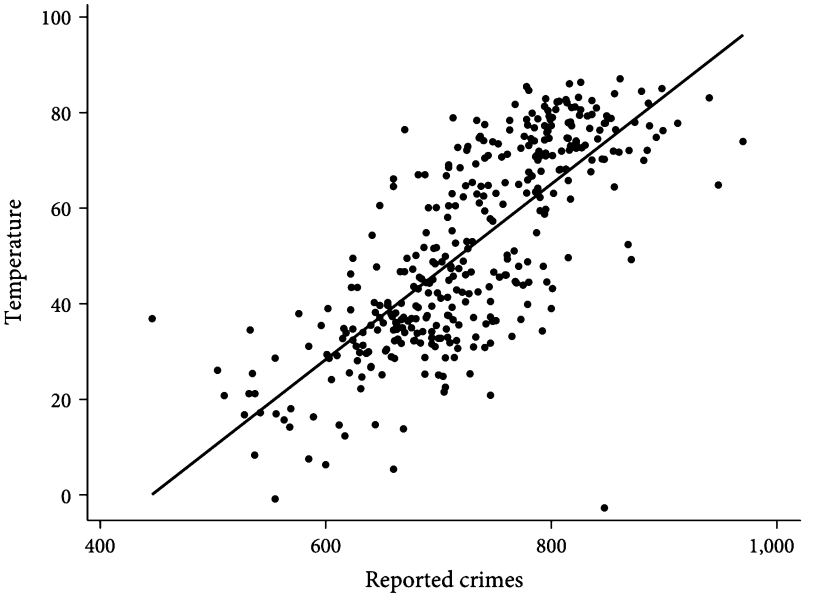

Figure 2.3. A line of best fit summarizing the relationship between temperature and crime in Chicago across days in 2018.

reported crime, on average, the temperature is 0.18 degrees higher. The sign of the slope (positive or negative) is the same regardless of which variable is on the horizontal or vertical axis because changing which variable is on which axis does not change whether they are positively or negatively correlated. But the actual number describing the slope and its substantive interpretation—that is, what it says about the world—has changed.

Fact or Correlation?

In order to establish whether a correlation exists, you must always make a comparison of some kind. For example, to learn about the correlation between temperature and crime, we need to compare hot and cold days and see whether the levels of crime differ, or alternatively, we can compare high- and low-crime days to see if their temperatures differ. This means that to assess the correlation between two variables, we need to have variation in both variables. For example, if we collected data only on days when the average temperature was 0 degrees, we would have no way of assessing the correlation between temperature and crime. And the same is true if we only examined days with five hundred reported crimes.

With this in mind, let's pause to check how clearly you are thinking about what a correlation is and how we learn about one. Don't worry if you aren't all the way there yet. Understanding whether a correlation exists turns out to be tricky. We are going to spend all of chapter 4 on this topic. Nonetheless, it is helpful to do a preliminary check now. So let's give it a try.

Think about the following statements. Which ones describe a correlation, and which ones do not?

1. People who live to be 100 years old typically take vitamins.
2. Cities with more crime tend to hire more police officers.
3. Successful people have spent at least ten thousand hours honing their craft.
4. Most politicians facing a scandal win reelection.
5. Older people vote more than younger people.

While each of these statements reports a fact, not all of those facts describe a correlation—that is, evidence on whether two features of the world tend to occur together. In particular, statements 1, 3, and 4 do not describe correlations, while statements 2 and 5 do. Let's unpack this.

Statements 1, 3, and 4 are facts. They come from data. They sound scientific. And if we added specific numbers to these statements, we could call them *statistics*. But not all facts or statistics describe correlations. The key issue is that these statements do not describe whether or not two features of the world tend to occur together—that is, they do not compare across different values of both features of the world.

To get a sense of this, focus on statement 4:

Most politicians facing a scandal win reelection.

Two features of the world are discussed. The first is whether a politician is facing a scandal. The second is whether the politician successfully wins reelection. The correlation being hinted at is a positive correlation between facing a scandal and winning reelection. But we don't actually learn from this statement of fact whether those two features of the world tend to occur together—that is, we have not compared the rate of reelection for those facing scandal to the rate of reelection for those not facing scandal.

We can assess this correlation, but not with the data described in statement 4. To assess the correlation, we'd need variation in both variables—facing a scandal and winning reelection. Just for fun, let's examine this correlation in some real data on incumbent members of the U.S. House of Representatives seeking reelection between 2006 and 2012. Scott Basinger from the University of Houston has systematically collected data on congressional scandals. Utilizing his data, let's see how many cases fall into four relevant cases: members facing a scandal who were reelected, members facing a scandal who were not reelected, scandal-free members who were reelected, and scandal-free members who were not reelected.

In table 2.2, we see that statement 4 is indeed a fact: 62 out of 70 (about 89%) members of Congress facing a scandal who sought reelection won. But we also see that most members of Congress not facing a scandal won reelection. In fact, 1,192 out of 1,293 (about 92%) of these scandal-free members won reelection. By comparing the scandal-plagued members to the scandal-free members, we now see that there is actually a slight negative correlation between facing a scandal and winning reelection.

Table 2.2. Most members of Congress facing a scandal are reelected, but scandal and reelection are negatively correlated.

	No Scandal	Scandal	Total
Not Reelected	101	8	109
Reelected	1,192	62	1,254
Total	1,293	70	1,363

We hope it is now clear why statement 4 does not convey enough information to know whether or not there is a correlation between scandal and reelection. The problem is that the statement is only about politicians facing scandal. It tells us that more of those politicians win reelection than lose. But to figure out if there is a correlation between scandal and winning reelection, we need to compare the share of politicians facing a scandal who win reelection to the share of scandal-free politicians who win. Had only 85 percent of the scandal-free members of Congress won reelection, there would be a positive correlation between scandal and reelection. Had 89 percent of them won, there would have been no correlation. But since we now know the true rate of reelection for scandal-free members was 92 percent, we see that there is a negative correlation. A similar analysis would show that statements 1 and 3 also don't convey enough information, on their own, to assess a correlation.

Statements 2 and 5 do describe correlations. Both statements make a comparison. Statement 2 tells us that cities with more crime have, on average, larger police forces than cities with less crime. And statement 5 tells us that older people tend to vote at higher rates than younger people. In both cases, we are comparing differences in one variable (police force size or voting rates) across differences in the other variable (crime rates or age). This is the kind of information you need to establish a correlation.

As we said at the outset, don't worry if you feel confused. Thinking clearly about what kind of information is necessary to establish a correlation, as opposed to just a fact, is tricky. We are going to spend chapter 4 making sure you really get it.

What Is a Correlation Good For?

Now that we have a shared understanding of what a correlation is, let's talk about what a correlation is good for. We've noted that correlations are perhaps the most important tool of quantitative analysts. But why? Broadly speaking, it's because correlations tell us what we should predict about some feature of the world given what we know about other features of the world.

There are at least three uses for this kind of knowledge: (1) description, (2) forecasting, and (3) causal inference. Any time we make use of a correlation, we want to think clearly about which of these three tasks we're attempting and what has to be true about the world for a correlation to be useful for that task in our particular setting.

Description

Describing the relationships between features of the world is the most straightforward use for correlations.

Why might we want to describe the relationship between features of the world? Suppose you were interested in whether younger people are underrepresented at the polls in a particular election, relative to their size in the population. A description of the relationship between age and voting might be helpful. Figure 2.4 shows a scatter plot of data on age and average voter turnout for the 2014 U.S. congressional election. In this figure, an observation is an age cohort. For each year of age, the figure shows the proportion of eligible voters who turned out to vote.

The figure also plots the line that best fits the data. This line has a slope of 0.006. In other words, on average, for every additional year of age, the chances that an individual turned out to vote in 2014 increases by 0.6 percentage points. So younger people do indeed appear to be underrepresented, as they turn out at lower rates than older people.

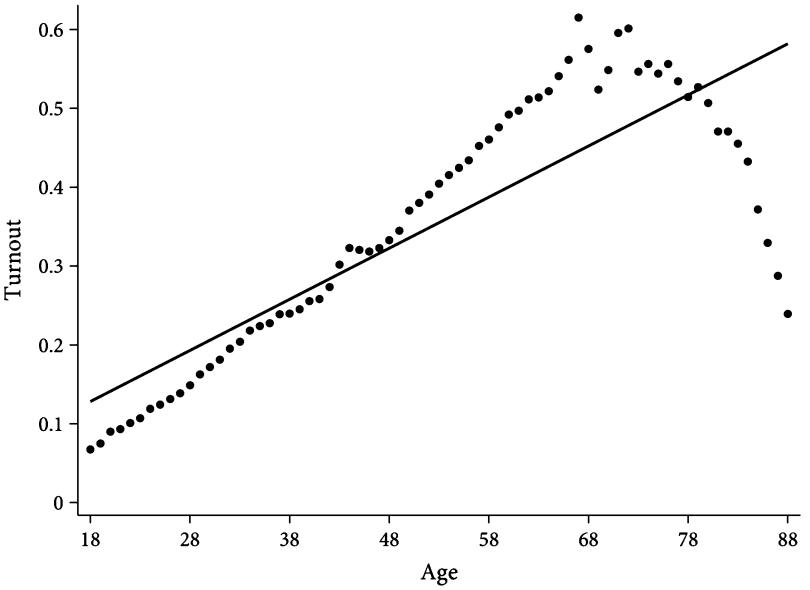

Figure 2.4. Voter turnout and age in the 2014 election.

This kind of descriptive analysis may be interesting in and of itself. It's important to know that younger people were less likely than older people to vote in 2014 and were therefore underrepresented in the electoral process. That relationship may inform how you think about the outcome of that election. Moreover, knowledge of this correlation might motivate you to further investigate the causes and consequences of the phenomenon of younger people turning out at low rates.

Of course, this descriptive relationship need not imply that these younger people will continue to vote at lower rates in future elections. So you can't necessarily use this knowledge to forecast future voter turnout. And it also doesn't mean that these younger people will necessarily become more likely to vote as they age. So you probably can't interpret this relationship causally. This descriptive analysis just tells us that older people were more likely to vote than younger people, on average, in the 2014 election. To push the interpretation further, you'd need to be willing to make stronger assumptions about the world, which we will now explore.

Forecasting

Another motivation for looking at correlations is *forecasting* or *prediction*—two terms that we will use interchangeably. Forecasting involves using information from some sample population to make predictions about a different population.

For instance, you might be using data on voters from past elections to make predictions about voters in future elections. Or you might be using the voters in one state to make predictions about voters in another state. Suppose you're running an electoral campaign, you have limited resources, and you're trying to figure out which of your supporters you should target with a knock on the door reminding them to turn out to vote. If you were already highly confident that an individual was going to vote in the absence

of your intervention, you wouldn't want to waste your volunteers' time by knocking on that door. So accurate forecasting of voter turnout rates could improve the efficiency of your campaign.

Correlations like the one above regarding age and voter turnout could be useful for this kind of forecasting. Since age is strongly correlated with turnout, it might be a useful variable for forecasting who is and is not already likely to vote. For instance, if you were able to predict, on the basis of age, that some group of voters is virtually certain to turn out even without your campaign efforts, you might want to focus your mobilization resources on other voters.

To use the correlation between age and voter turnout for forecasting in this way, you don't need to know why they are correlated. But, unlike if you just want to *describe* the relationship between age and voter turnout in the 2014 election, if you want to *forecast*, you need to be willing to make some additional assumptions about the world.

This raises two important concerns that you must think clearly about in order to use correlation for forecasting responsibly. The first is whether the relationship you found in your sample is indicative of a broader phenomenon or whether it is the result of chance variation in your data. Answering this question requires *statistical inference*, which is the topic of chapter 6. Second, even if you are convinced that you've found a real relationship in your sample, you'll want to think about whether your sample is representative of the population about which you are trying to make predictions. We will explore representativeness in greater detail in our discussion of samples and external validity in chapters 6 and 16.

Let's go back to using age and voter turnout from one election to make predictions about another election. Doing so only makes sense if it is reasonable to assume that the relationship between these two variables isn't changing too quickly. That is, the correlation between age and voter turnout in, for example, the 2014 election would only be useful for figuring out which voters to target in the 2016 election if it seems likely that the relationship between age and turnout in 2016 will be more or less the same as the relationship between age and turnout in 2014. Similarly, if you only had data on age and voter turnout in the 2014 election for twenty-five states, you might use the correlation between age and turnout in those states to inform a strategy in the other twenty-five states. But this would only be sensible if you had reason to believe that the relationship between age and turnout was likely to be similar in the states on which you did and did not have data.

You'd also want to take care in making predictions beyond the range of available data. Our data tell us voter turnout rates for voters ages 18–88. Lines, however, go on forever. So the line of best fit gives us predictions for any age. But we should be careful extrapolating our predictions about voter turnout to, say, 100-year-olds, since we don't have any data for them, so we can't know whether the relationship described by the line is likely to hold for them or not, even for the 2014 election. And we can be sure the line's predictions for turnout by 10-year-olds won't be accurate—they aren't even allowed to vote.

Relatedly, when using some statistic, like the slope of a line of best fit, to do prediction, we need to think about whether the relationship is actually linear. If not, a linear summary of the relationship might be misleading. We'll discuss this in greater detail below.

It is worth noting that, in practical applications, it would be unusual to try to do forecasting simply using the correlation between two variables. One might, instead, try to predict voter turnout using its relationship with a host of variables like gender, race,

income, education, and previous voter turnout. We'll discuss such multivariable and conditional correlations in chapter 5.

Using data for forecasting and prediction is a rapidly growing area for analysts in policy, business, policing, sports, government, intelligence, and many other fields. For instance, suppose you're running your city's public health department. Every time you send a health inspector to a restaurant, it costs time and money. But restaurant violations of the health code do harm to your city's residents. Therefore, you would very much like to send inspectors to those restaurants that are most likely to be in violation of the health codes, so as not to waste time and money on inspections that don't end up improving public safety. The more accurately you can forecast which restaurants are in violation, the more effectively you can deploy your inspectors. You could imagine using data on restaurants that did and did not violate health codes in the past to try to predict such violations on the basis of their correlation with other observable features of a restaurant. Plausibly useful restaurant features might include Yelp reviews, information about hospital visits for food poisoning, location, prices, and so on. Then, with these correlations in hand, you could use future Yelp reviews and other information to predict which restaurants are likely in violation of the health codes and target those restaurants for inspection.

This example points to another tricky issue. The very act of using correlations for prediction can sometimes make correlations that held in the past cease to hold in the future. For instance, suppose the health department observes a strong correlation between restaurants that are open twenty-four hours a day and health code violations. On the basis of that correlation, they might start sending health inspectors disproportionately to twenty-four-hour restaurants. A savvy restaurant owner who becomes aware of the new policy might adapt to fool the health department, say closing from 2:00 to 3:00 a.m. every night. This small change in operating hours would presumably do nothing to clean up the restaurant. But the manager would have gamed the system, rendering predictions based on past data inaccurate for the future. We'll discuss this general problem of adaptation in greater detail in chapter 16.

Forecasting would also be useful to a policy maker who would like to know the expected length of an economic downturn for budgetary purposes, a banker who wants to know the credit worthiness of potential borrowers, or an insurance company that wants to know how many car accidents a potential client is likely to get in this year. The managers of our beloved Chicago Bears would love to predict which college football players could be drafted to increase the team's chances of winning a Super Bowl. But given their past track record, we don't hold out much hope. Data can't work miracles.

It is also worth thinking about the potential ethical implications of using predictions to guide behavior. For instance, research finds that consumer complaints about cleanliness in online restaurant reviews are positively correlated with health code violations. This is potentially useful predictive information—governments could use data collected from review sites to figure out where to send restaurant inspectors. In response to such findings, an article in *The Atlantic* declared, "Yelp might clean up the restaurant industry." But a study by Kristen Altenburger and Daniel Ho shows that online reviewers are biased against Asian restaurants—comparing restaurants that received the same score from food-safety inspectors, they find that reviewers were more likely to complain about cleanliness in the Asian restaurants. This means that if governments make use of the helpful predictive correlation between online reviews and health code violations, it will inadvertently discriminate against Asian restaurants by disproportionately targeting them for inspection. Do you want your government to make use of such

information? Or are there ethical or social costs of targeting restaurants for inspection in an ethnically biased way that outweigh the benefits of more accurate predictions? We will return to some of these ethical issues at the end of the book.

Causal Inference

Another reason we might be interested in correlations is to learn about causal relationships. Many of the most interesting questions that quantitative analysts face are inherently causal. That is, they are about how changing some feature of the world would cause a change in some other feature of the world. Would lowering the cost of college improve income inequality? Would implementing a universal basic income reduce homelessness? Would a new marketing strategy boost profits? These are all causal questions. As we'll see throughout the book, using correlations to make inferences about causal relationships is common. But it is also fraught with opportunities for unclear thinking. (Understanding causality will be the subject of the next chapter.)

Using correlation for causal inference has all the potential issues we just discussed when thinking about using correlation for prediction and there are new issues. The key one is that correlation need not imply causation. That is, a correlation between two features of the world doesn't mean one of them causes the other.

Suppose you want to know the effect of high school math training on subsequent success in college. This is an important question if you're a high school student, a parent or counselor of a high school student, or a policy maker setting educational standards. Will high school students be more likely to attend and complete college if they take advanced math in high school?

As it turns out, the correlation between taking advanced math and completing college is positive and quite strong—for instance, people who take calculus in high school are much more likely to graduate from college than people who do not. And the correlation is even stronger for algebra 2, trigonometry, and pre-calculus. But that doesn't mean that taking calculus causes students to complete college.

Of course, one possible source of this correlation is that calculus prepares students for college and causes them to become more likely to graduate. But that isn't the only possible source of this correlation. For instance, maybe, on average, kids who take calculus are more academically motivated than kids who don't. And maybe motivated kids are more likely to complete college regardless of whether or not they take calculus in high school. If that is the case, we would see a positive correlation between taking calculus and completing college even if calculus itself has no effect on college completion. Rather, whether a student took calculus would simply be an indirect measure of motivation, which is correlated with completing college.

What's at stake here? Well, if the causal story is right, then requiring a student to take calculus who otherwise wouldn't will help that student complete college by offering better preparation. But if the motivation story is right, then requiring that student to take calculus will not help with college completion. In that story, calculus is just an indicator of motivation. Requiring a student to take calculus does not magically make that student more motivated. It could even turn out that requiring that student to take calculus might impose real costs—in terms of self-esteem, motivation, or time spent on other activities—without any offsetting benefits.

The exact mistake we just described was made in a peer-reviewed scientific article. The researchers compared the college performance of people who did and did not take a variety of intensive high school math courses. On the basis of a positive correlation, they

suggested that high school counselors "use the results of this study to inform students and their parents and guardians of the important role that high school math courses play with regard to subsequent bachelor's degree completion." That is, they mistook correlation for causation. On the basis of these correlations, they recommended that students who were not otherwise planning to do so should enroll in intensive math courses to increase their chances of graduating from college.

We'll return to the problem of mistaking correlation for causation in part 3. For now, you should note that, although purported experts do it all the time, in general, it is wrong to infer causality from correlations.

Measuring Correlations

There are several common statistics that can be used to describe and measure the correlation between variables. Here we discuss three of them: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*. But before going through these three different ways of measuring correlations, we need to talk about means, variances, and standard deviations—statistics that help us summarize and understand variables.

Mean, Variance, and Standard Deviation

Let's focus on our Chicago crime and temperature data. Recall that in this data set, each observation is a day in 2018. And for each day we observe two variables, the number of reported crimes and the average temperature as measured in degrees Fahrenheit at Midway Airport. We aren't going to reproduce the entire data set here, since it has 365 rows (one for each day of 2018). Table 2.3 shows what the data looks like for the month of January. For the remainder of this discussion, we will treat the days of January 2018 as our population of interest.

For any observation i, call the value of the crime variable $crime_i$ and the value of the temperature variable $temperature_i$. In our data table, i can take any value from 1 through 31, corresponding to the thirty-one days of January 2018. So, for instance, the temperature on January 13 was $temperature_{13} = 12.3$, and the number of crimes reported on January 24 was $crime_{24} = 610$.

A variable has a *distribution*—a description of the frequency with which it takes different values. We often want to be able to summarize a variable's distribution with a few key statistics. Here we talk about three of them.

It will help to have a little bit of notation. The symbol $\sum$ (the upper-case Greek letter *sigma*) denotes summation. For example, $\sum_{i=1}^{31} \text{crime}_i$ is the sum of all the values of the crime variable from day 1 through day 31. To find it, you take the values of crime for day 1, day 2, day 3, and so on through 31 and sum (add) them together. That is, you add up $\text{crime}_1 = 847$ and $\text{crime}_2 = 555$ and $\text{crime}_3 = 568$ and so on through $\text{crime}_{31} = 708$. You find these specific values for the crime variable on each day by referring back to the data in table 2.3.

Now we can calculate the *mean* of each variable's distribution. (Sometimes this is just called the *mean of the variable*, leaving reference to the distribution implicit). The mean is denoted by μ (the Greek letter *mu*). The mean is just the average. We find it by summing the values of the observations (which we now have convenient notation for) and dividing by the number of observations. For January 2018, the means of our two variables are

Table 2.3. Average temperature at Chicago Midway Airport and number of crimes reported in Chicago for each day of January 2018.

Day	Temperature (°F)	Crimes
1	−2.7	847
2	−0.9	555
3	14.2	568
4	6.3	600
5	5.4	660
6	7.5	585
7	25.4	535
8	33.9	618
9	30.1	653
10	44.9	709
11	51.7	698
12	21.6	705
13	12.3	617
14	15.7	563
15	16.8	528
16	14.6	612
17	14.7	644
18	25.6	621
19	34.8	707
20	40.4	724
21	42.9	716
22	48.9	722
23	32.3	716
24	29.2	610
25	35.5	640
26	46.0	759
27	45.6	754
28	35.0	668
29	25.2	650
30	24.7	632
31	37.6	708
Mean	26.3	655.6
Variance	220.3	5183.0
Standard deviation	14.8	72.0

$$\mu_{\text{crime}} = \frac{\sum_{i=1}^{31} \text{crime}_i}{31} = \frac{847 + 555 + \cdots + 708}{31} = 655.6$$

and

$$\mu_{\text{temperature}} = \frac{\sum_{i=1}^{31} \text{temperature}_i}{31} = \frac{-2.7 + -0.9 + \cdots + 37.6}{31} = 26.3.$$

A second statistic of interest is the *variance*, which we denote by σ^2 (the lower-case Greek letter *sigma*, squared). We'll see why it is squared in a moment. The variance is a way of measuring how far from the mean the individual values of the variable tend to be. You might even say that the variance measures how variable the variable is. (You can also think of it, roughly, as a measure of how spread out the variable's distribution is.)

Here's how we calculate the variance. Suppose we have some variable X (like crime or temperature). For each observation, calculate the *deviation* of that observation's value of X from the mean of X. So, for observation i, the deviation is the value of X for observation i (X_i) minus the mean value of X across all observations (μ_X)—that is, $X_i - \mu_X$. On January 13, 2018, the temperature was 12.3 degrees Fahrenheit. The mean temperature in January 2018 was 26.3 degrees Fahrenheit. So January 13's deviation from the January mean was $12.3 - 26.3 = -14$. That is, January 13, 2018, was fourteen degrees colder than the average day in January 2018. By contrast, the deviation of January 23, 2018, was $32.3 - 26.3 = 6$. On January 23, it was six degrees warmer than on the average day in January 2018.

Note that these deviations can be positive or negative since observations can be larger or smaller than the mean. But for the purpose of measuring how variable the observations are, it doesn't matter whether any given deviation is positive or negative. We just want to know how far each observation is from the mean in any direction. So we need to transform the deviations into positive numbers that just measure the distance from the mean rather than the sign and distance. To do this, we could look at the absolute value of the deviations. But for reasons we'll discuss later, we typically make the deviations positive by squaring them instead. The variance is the average value of these squared deviations. So, if there are N observations (in our data, $N = 31$) the variance is

$$\sigma_X^2 = \frac{\sum_i^N (X_i - \mu_X)^2}{N}.$$

For the two variables in our data, the variances are

$$\begin{aligned}\sigma^2_{\text{crime}} &= \frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31} \\ &= \frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31} \approx 5183\end{aligned}$$

and

$$\begin{aligned}\sigma^2_{\text{temperature}} &= \frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31} \\ &= \frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31} \approx 220.3.\end{aligned}$$

By focusing on the average of the squared deviations rather than on the average of the absolute value of the deviations, the variance is putting more weight on observations that are farther from the mean. If the richest person in society gets richer, this increases the variance in wealth more than if a moderately rich person gets richer by the same amount. For example, suppose the average wealth is 1. If someone with a wealth of 10 gains 1 more unit of wealth, the variance increases by $\frac{10^2 - 9^2}{N} = \frac{19}{N}$. But if someone with a wealth of 100 gains one more unit of wealth, the variance increases by $\frac{100^2 - 99^2}{N} = \frac{199}{N}$.

The variance is a fine measure of how variable a variable is. But since we've squared everything, there is a sense in which it is not measured on the same scale as the variable itself. Sometimes we want a measure of variability that is on that same scale. When that is the case, we use the *standard deviation*, which is just the square root of the variance. We denote the standard deviation by σ (the lower-case Greek letter *sigma*):

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{\frac{\sum_i^N (X_i - \mu_X)^2}{N}}.$$

The standard deviation—which is also a measure of how spread out a variable's distribution is—roughly corresponds to how far we expect observations to be from the mean, on average. Though, as we've noted, compared to the average absolute value of the deviations, it puts extra weight on observations that are farther from the mean.

For the two variables in our data, the standard deviations are

$$\begin{aligned}\sigma_{\text{crime}} &= \sqrt{\frac{\sum_{i=1}^{31} (\text{crime}_i - \mu_{\text{crime}})^2}{31}} \\ &= \sqrt{\frac{(847 - 655.6)^2 + (555 - 655.6)^2 + \cdots + (708 - 655.6)^2}{31}} \approx 72\end{aligned}$$

and

$$\begin{aligned}\sigma_{\text{temperature}} &= \sqrt{\frac{\sum_{i=1}^{31} (\text{temperature}_i - \mu_{\text{temperature}})^2}{31}} \\ &= \sqrt{\frac{(-2.7 - 26.3)^2 + (-0.9 - 26.3)^2 + \cdots + (37.6 - 26.3)^2}{31}} \approx 15.1.\end{aligned}$$

Now that we understand what a mean, variance, and standard deviation are, we can discuss three important ways in which we measure correlations: the *covariance*, the *correlation coefficient*, and the *slope of the regression line*.

Covariance

Suppose we have two variables, like crime and temperature, and we want to measure the correlation between them. One way to do this would be to calculate their *covariance* (denoted *cov*). To keep our notation simple, let's call those two variables X and Y. And let's assume we have a population of size N.

Here's how you calculate the covariance. For every observation, calculate the deviations—that is, how far the value of X is from the mean of X and how far the value of Y is from the mean of Y. Now, for each observation, multiply the two deviations together, so you have $(X_i - \mu_X)(Y_i - \mu_Y)$ for each observation i. Call this the *product of the deviations*. Finally, to find the covariance of X and Y, calculate the average value of this product:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

Let's see that the covariance is a measure of the correlation. Consider a particularly strong version of positive correlation: suppose whenever X is bigger than average ($X_i - \mu_X > 0$), Y is also bigger than average ($Y_i - \mu_Y > 0$), and whenever X is smaller than average ($X_i - \mu_X < 0$), Y is also smaller than average ($Y_i - \mu_Y < 0$). In this case, the product of the deviations will be positive for every observation—either both deviations will be positive, or both deviations will be negative. So the covariance will be positive, reflecting the positive correlation. Now consider a particularly strong version of negative correlation: suppose whenever X is bigger than average, Y is smaller than average, and whenever X is smaller than average, Y is bigger than average. In this case, the product of the deviations will be negative for every observation—one deviation is always negative and the other always positive. So the covariance will be negative, reflecting the negative correlation. Of course, neither of these extreme cases has to hold. But if a larger-than-average X usually goes with a larger-than-average Y, then the covariance will be positive, reflecting a positive correlation. If a larger-than-average X usually goes with a smaller-than-average Y, then the covariance will be negative, reflecting a negative correlation. And if the values of X and Y are unrelated to each other, the covariance will be zero, reflecting the fact that the variables are uncorrelated.

Correlation Coefficient

While the meaning of the sign of the covariance is clear, its magnitude can be a bit hard to interpret, since the product of the deviations depends on how variable the variables are. We can get a more easily interpretable statistic that still measures the correlation by accounting for the variance of the variables.

The *correlation coefficient* (denoted *corr*) is simply the covariance divided by the product of the standard deviations:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

When we divide the covariance by the product of the standard deviations, we are normalizing things. That is, the covariance could, in principle, take any value. But the correlation coefficient always takes a value between -1 and 1. A value of 0 still indicates no correlation. A value of 1 indicates a positive correlation and perfect linear dependence—that is, if you made a scatter plot of the two variables, you could draw a straight, upward-sloping line through all the points. A value of -1 indicates a negative correlation and perfect linear dependence. A value between 0 and 1 indicates positive correlation but not a perfect linear relationship. And a value between -1 and 0 indicates negative correlation but not a perfect linear relationship.

The correlation coefficient is sometimes denoted by the letter r. And we also sometimes square the correlation coefficient to compute a statistic called r-squared or r^2. This statistic always lies between 0 and 1.

One potentially attractive feature of the r^2 statistic is that it can be interpreted as a proportion. It's often interpreted as the proportion of the variation in Y *explained* by X or, equivalently, the proportion of X explained by Y. As we'll discuss in later chapters, the word *explained* can be misleading here. It doesn't mean that the variation in X causes the variation in Y or vice versa. It also doesn't account for the possibility that this observed correlation might have arisen by chance rather than reflecting a genuine phenomenon in the world.

Slope of the Regression Line

One potential concern with the correlation coefficient and the r^2 statistic is that they don't tell you anything about the substantive importance or size of the relationship between X and Y. Suppose our two variables of interest are crime and temperature in Chicago. A correlation coefficient of .8 tells us that there is a strong, positive relationship between the two variables, but it doesn't tell us what that relationship is. It could be that every degree of temperature corresponds with .1 extra crimes, or it could be that every degree of temperature corresponds with 100 extra crimes. Both are possible with a correlation coefficient of .8. But they mean very different things.

For this reason, we don't spend much time thinking about these ways of measuring correlation. We typically focus on the slope of a line of best fit, as we've already shown you. Moreover, we tend to focus on one particular way of defining which line fits best. Remember, a line of best fit minimizes how far the data points are from the line on average. We typically measure how far a data point is from the line with the square of the distance from the data point to the line (so every value is positive, just like with squaring deviations). We focus on the line of best fit that minimizes the sum of these squared distances (or the *sum of squared errors*). This particular line of best fit is called the ordinary least squares (OLS) regression line, and usually, when someone just says *regression line*, they mean *OLS regression line*. All the lines of best fit we drew earlier in this chapter were OLS regression lines.

The slope of the regression line, it turns out, can be calculated from the covariance and variance. The slope of the regression line (also sometimes called the *regression coefficient*) when Y is on the vertical axis and X is on the horizontal axis is

$$\frac{\text{cov}(X, Y)}{\sigma_X^2}.$$

This number tells us, descriptively, how much Y changes, on average, as X increases by one unit. Had we divided by σ_Y^2 instead of σ_X^2, then it would tell us how much X changes, on average, as Y increases by one unit. As we've seen, those can be different numbers.

We'll spend a lot more time on regression lines in chapters 5 and 10.

Populations and Samples

Before moving on, there is one last issue that is worth pausing to highlight. We can think about each of the statistics we've talked about—the mean, the variance, the covariance, the correlation coefficient, the slope of the regression line—in two ways. There

is a value of each of those statistics that corresponds to the whole population we are interested in. And there is a value of those statistics that corresponds to the sample of data we might happen to have. Either value can be of interest, but they can be importantly different. We have avoided that issue here by focusing on a case where our data and our population are the same—we have crime and temperature for every day in January 2018, which we've treated as our population and our sample. But this won't always be the case. For instance, we might have been interested in the relationship between crime and temperature in January over many years but only had a sample of data for the year 2018. This would give rise to all sorts of questions about what we can learn about January 2019 or January 1918 from our 2018 data. We will revisit these issues in chapter 6.

Straight Talk about Linearity

All of the various ways of measuring correlations that we have discussed focus on assessing linear relationships between variables. We will delve into this topic in more detail later on, especially in chapter 5 when we return to the topic of age and voter turnout in the context of our discussion of regression. But for now we will note that linear relationships are often interesting and important, but not all interesting and important relationships are linear. Consider, for example, the two possible relationships between the variables X and Y illustrated in figure 2.5.

As the regression lines make clear, in both these figures, the correlation between X and Y is 0. But these relationships are clearly different, just not in a way that is captured by the regression line.

In the left panel, there is no correlation between X and Y and there also doesn't seem to be any interesting relationship of any kind. You really can't predict the value of Y from X or vice versa. In the right panel, there is also no correlation between X and Y—on average, high values of X don't tend to occur with high values of Y, nor do low values of X tend to occur with low values of Y. But there is certainly a relationship between these two variables. In fact, X is quite useful in predicting Y in the right panel. This teaches us a lesson. Clear thinking about data requires more than just computing correlations. Among other things, it is important to look at your data (e.g., with scatter plots like these), lest you miss interesting nonlinear relationships.

There are lots of statistical approaches for dealing with non-linearity, and we'll discuss some of them in this book. But, as it turns out, linear tools for describing data can still be useful, even when the variables are related in a non-linear way. For instance, in the right panel of figure 2.5, there is a strong negative correlation between X and Y when X is less than 0 and a strong positive correlation between X and Y when X is greater than 0. So one thing we could do with linear tools is draw two lines of best fit, one for when X is less than 0 and one for when it is greater than 0. That would look like figure 2.6.

Another thing we could do is transform one of the variables so that the relationship looks more linear. For instance, in our example, although there is no correlation between Y and X, there is a strong linear relationship between Y and X^2. In figure 2.7 we plot X^2 on the horizontal axis and Y on the vertical axis. When we transform X into X^2, negative values of X become positive values of X^2 (e.g., -1 becomes 1), while the positive values stay positive (e.g., 1 stays 1). So it is as if we are folding the figure in

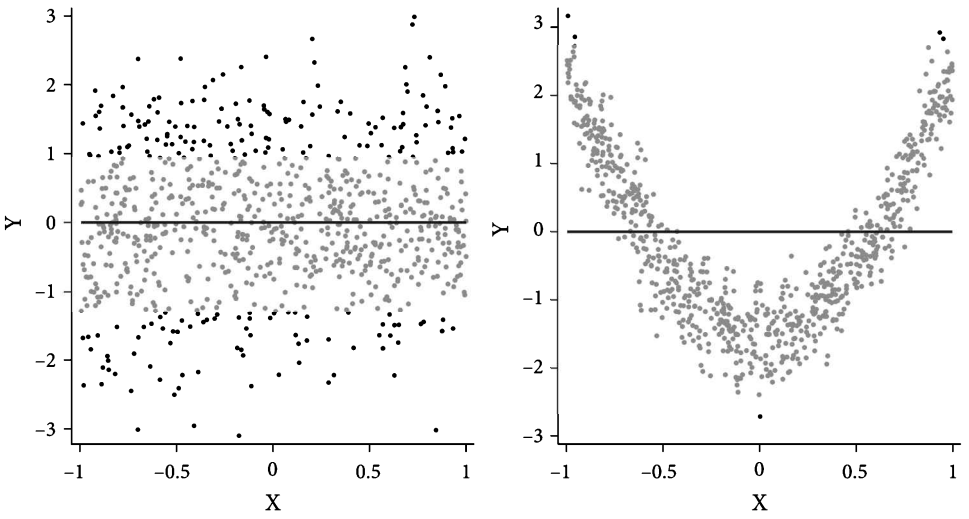

Figure 2.5. Zero correlation can mean many things.

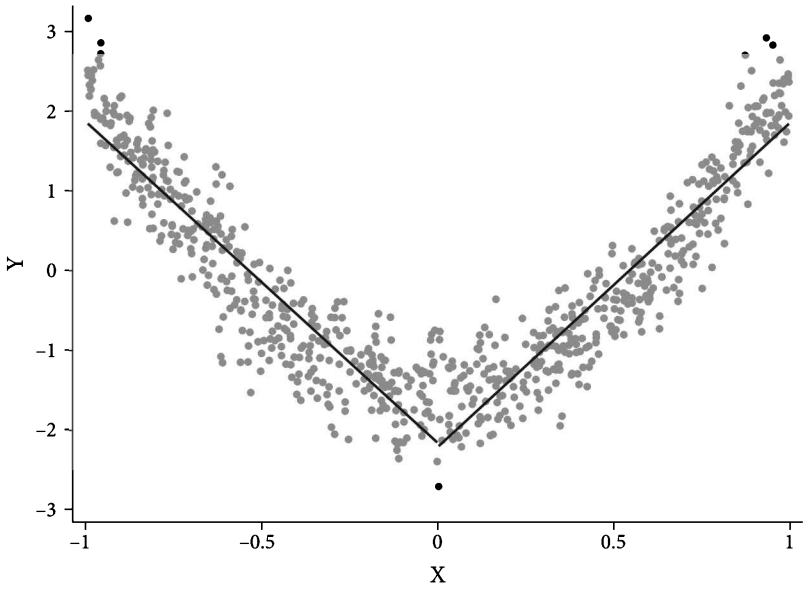

Figure 2.6. Fitting two separate regression lines to a non-linear relationship.

on itself at $X = 0$, and then we're twisting and stretching it a little so that X becomes X^2 (0 stays at 0, 1 stays at 1, .5 becomes $.5^2 = .25$, and so on).

With this transformation, our regression line shows a strong positive relationship between Y and X^2, and we can do a good job describing the relationship between these variables with our linear tools.

It's also worth pointing out that describing the relationship between two variables with a linear function is always appropriate when we're dealing with binary variables.

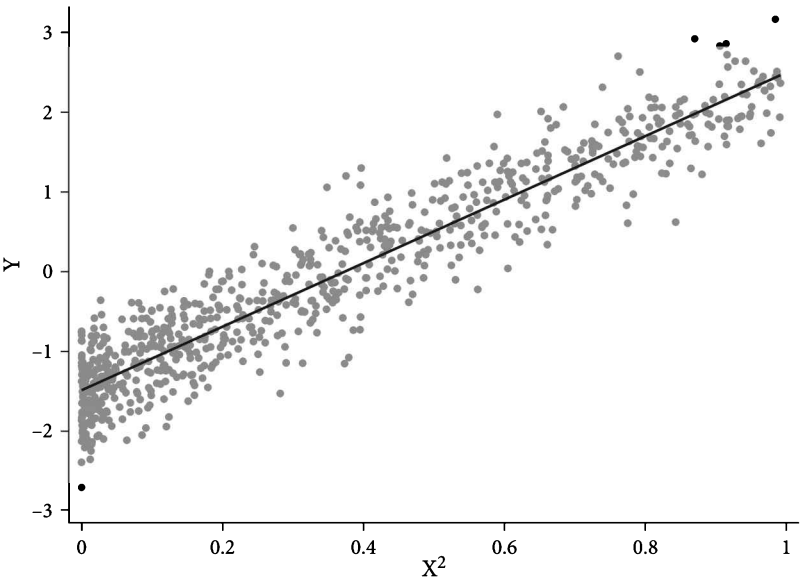

Figure 2.7. Creating a linear relationship by transforming a variable.

For example, let's return to the correlation between oil production and autocracy. Figure 2.8 plots the data. The scatter plot is not very interesting or informative because there are only four possible combinations of our two variables. Accordingly, all of the data points lie on one of those four dots (although we have attempted to make the scatter plot more informative by making the size of the dots proportional to the number of countries at each set of values). However, we can still plot the slope of the regression line. The slope of this line is simply the proportion of major oil-producing countries that are autocracies minus the proportion of non-major oil-producing countries that are autocracies. In other words, we learn the same thing from this picture that we learned from the table at the outset of the chapter.

One reason that we focus so much on linear relationships is that even non-linear relationships tend to look approximately linear if you zoom in enough—that is, if you are interested in a sufficiently small range of values of the variable X. We must be particularly cautious about extrapolating when we zoom in like that. As we move farther from the range of data in which the relationship is approximately linear, our descriptions of the relationship (and, by extension, any predictions we make) will be less and less accurate.

To think more about the dangers of extrapolation, consider an example. Political analysts find that the incumbent party in U.S. presidential elections tends to get about 46 percent of the vote when there is 0 income growth, and an extra 3.5 percentage points of the vote for every percentage point increase in income growth. Of course, they've measured this relationship using data on income growth levels that have actually occurred. Does this mean that we should predict incumbent vote share will be 81 percent if income growth is 10 percent? Probably not. And the incumbent's vote share definitely would not be 116 percent if income growth were 20 percent—that's impossible! But that doesn't mean a linear description of the data isn't useful for the range of income growths that we actually experience.

Figure 2.8. A regression line through data with a binary variable gives the difference in means.

Wrapping Up

Correlations form the foundation of data analysis. They are the way we talk about relationships between features of the world. And the various statistics by which we measure correlations—like the covariance, correlation coefficient, or slope of the regression line—are the way we quantify those relationships.

As we've discussed, correlations can be used for a variety of purposes including description, forecasting, and causal inference. In chapter 3, we turn our focus to causality in order to understand what it means and start to get a handle on the aphorism with which we began—correlation need not imply causation. However, a fuller understanding of the relationship between correlation and causation will have to wait until chapter 9.

Key Terms

- **Correlation:** The correlation between two features of the world is the extent to which they tend to occur together.
- **Positively correlated:** When higher (lower) values of one variable tend to occur with higher (lower) values of another variable, we say that the two variables are positively correlated.
- **Negatively correlated:** When higher (lower) values of one variable tend to occur with lower (higher) values of another variable, we say that the two variables are negatively correlated.
- **Uncorrelated:** When there is no correlation between two variables, meaning that higher (lower) values of one variable do not systematically coincide with higher or lower values of the other variable, we say that they are uncorrelated.
- **Line of best fit:** A line that minimizes how far data points are from the line on average, according to some measure of distance from data to the line.

- **Mean (μ)**: The average value of a variable.
- **Deviation from the mean:** The distance between an observation's value for some variable and the mean of that variable.
- **Variance (σ^2):** A measure of how variable a variable is. It is the average of the square of the deviations from the mean.
- **Standard deviation (σ):** Another measure of how variable a variable is. The standard deviation is the square root of the variance. It has the advantage of being measured on the same scale as the variable itself and roughly corresponds to how far the typical observation is from the mean (though, like the variance, it puts more weight on observations far from the mean).
- **Covariance (cov):** A measure of the correlation between two variables. It is calculated as the average of the product of the deviations from the mean.
- **Correlation coefficient (r):** Another measure of the correlation between two variables. It is calculated as the covariance divided by the product of the variances. The correlation coefficient takes a value between -1 and 1, with -1 reflecting perfect linear negative dependence, 0 reflecting no correlation, and 1 reflecting perfect linear dependence.
- **r^2:** The square of the correlation coefficient. It takes values between 0 and 1 and is often interpreted as the proportion of the variation in one variable explained by the other variable. But we have to pay careful attention to what we mean by "explained." Importantly, it doesn't mean that variation in one variable causes variation in the other.
- **Sum of squared errors:** The sum of the square of the distance from each data point to a given line of best fit. This gives us one way of measuring how well the line fits/describes/explains the data.
- **OLS regression line:** The line that best fits the data, where *best fits* means that it minimizes the sum of squared error.
- **Slope of a line:** The slope of a line tells you how much the line changes on the vertical axis as you move one unit along the horizontal axis. So a completely horizontal line has a slope of 0. An upward sloping 45-degree line has a slope 1, a downward sloping 45-degree line has a slope of -1, and so on.
- **Slope of the regression line or regression coefficent:** The slope of the regression line describes how the value of one variable changes, on average, when the other variable changes. The slope of the regression line is the covariance of two variables divided by the variance of one of them, sometimes also called the regression coefficient.

Exercises

2.1 Consider the following three statements. Which ones describe a correlation, and which ones do not? Why?

(a) Most professional data analysts took a statistics course in college.
(b) Among Major League Baseball players, pitchers tend to have lower-than-average batting averages. (We'll learn why this is the case in chapter 16.)
(c) Whichever presidential candidate wins Ohio tends to win the Electoral College.

2.2 Consider the last statement about Ohio and presidential elections. Do you think it's useful for description? Forecasting? Causal inference? Why or why not?

2.3 The table below shows some data on which countries are major oil producers and which countries experienced a civil war between 1946 and 2004. Are being a major oil producer and experiencing civil war positively correlated, negatively correlated, or uncorrelated? Explain your answer.

	Civil War	**No Civil War**
Oil Producer	7	12
Non-Oil Producer	55	94

2.4 The table below provides data about height and income among American men, taken from the National Longitudinal Survey. It is fine to use a calculator for this question, but don't use a spread sheet or statistical software to compute the answers.

Height (in)	**Average Income $**
60	39,428
61	35,087
62	40,575
63	39,825
64	55,508
65	56,377
66	59,746
67	66,699
68	59,787
69	66,176
70	79,202
71	70,432
72	77,975
73	72,606
74	71,063
75	80,330

(a) Calculate the mean of each of these variables.
(b) Calculate the variance of each of these variables.
(c) Calculate the standard deviation of each of these variables.
(d) Calculate the covariance between these two variables.
(e) Calculate the correlation coefficient for these variables.
(f) Are the two variables positively correlated, negatively correlated, or uncorrelated? Explain your answer.

Readings and References

For more on the corruption data we discussed take a look at

Scott J. Basinger. 2013. "Scandals and Congressional Elections in the Post-Watergate Era." *Political Research Quarterly* 66(2):385–398.

For more information about the Polity IV Project, which classifies countries as democratic or autocratic, see https://www.systemicpeace.org/polity/polity4.htm.

We discussed two articles on using online reviews to predict health code violations:

Emily Badger. 2013. "How Yelp Might Clean Up the Restaurant Industry." *The Atlantic*. July/August.

Kristen M. Altenburger and Daniel E. Ho. 2018. "When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions." *Journal of Institutional and Theoretical Economics* 174(1):98–122.

The study of advanced math and college completion is.

Jerry Trusty and Spencer G. Niles. 2003. "High-School Math Courses and Completion of the Bachelor's Degree." *Professional School Counseling* 7(2):99–107.

If you are interested in examples of the growing use of forecasting and prediction in addressing important policy problems, have a look at

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5):491–95.

CHAPTER 3

Causation: What Is It and What Is It Good For?

What You'll Learn

- A causal effect is a change in some feature of the world that would result from a change to some other feature of the world.
- Assessing causal relationships is crucial for policy and decision making.
- *"What effect did this have on the outcome?"* is a more conceptually clear question than *"What caused the outcome?"*
- Causal relationships are about comparisons of *counterfactual* worlds. As a result, they are fundamentally unobservable. But, in certain situations, we can learn about them from data.

Introduction

As we saw in chapter 2, knowledge of correlations is useful for many purposes. Among the most important, but also most vexing, purposes is learning about causal relationships.

We make claims about causal knowledge all the time. I did poorly on the test because I didn't get enough sleep. Going to college will improve my future job prospects. A political candidate lost an election because of an attack ad. Violent crime is down because of a new policing strategy.

Thinking clearly about whether a causal relationship exists is perhaps the most important conceptual challenge for learning to use information to make better decisions. This is because causal knowledge is the key to understanding how your decisions and actions affect the world around you. If you propose a new tax policy, test-prep strategy, exercise plan, or advertising campaign, you're doing so not because you think it is correlated with better outcomes. Rather, you must believe that enacting your proposal will actually cause better outcomes.

Our goal in this chapter is to clarify exactly what we mean when we talk about causal relationships. Causality is a deep and perplexing topic to which much attention has been paid by scholars from many different fields. We won't be able to resolve all the thorny philosophical questions here. Instead we've set more modest goals. First, we want to make sure we are all on the same page by defining how we will use causal language for the duration of this book. Then we will explain why the notion of causality we adopt is a particularly valuable one. Finally, we will discuss some other approaches to talking

about causality and explain why, from our point of view, they are less helpful than the one we adopt.

What Is Causation?

When we talk about causation, we're talking about the effect of one thing on another. In non-technical terms, a *causal effect* is a change in some feature of the world that would result from a change to some other feature of the world. So, for instance, we would say that the tax rate has a causal effect on government revenue if changing the tax rate would lead to a change in government revenue.

We've defined the notion of an effect in non-technical terms, so you might not have noticed that we actually slipped in a bit of philosophy. What do we mean by *would result*? After all, the world is as it is. Where did this change in some other feature of the world come from?

That's a good question. In fact, our definition of a causal effect relies on a thought experiment about which we need to be explicit. Let's start with an example.

The movie star Gwyneth Paltrow runs a company called Goop that promotes stickers, called Body Vibes, that are supposed to promote health, wellness, *and* good skin. Here's what the Goop website says about Body Vibes:

> Human bodies operate at an ideal energetic frequency, but everyday stresses and anxiety can throw off our internal balance, depleting our energy reserves and weakening our immune systems. Body Vibes stickers come pre-programmed to an ideal frequency, allowing them to target imbalances. While you're wearing them—close to your heart, on your left shoulder or arm—they'll fill in the deficiencies in your reserves, creating a calming effect, smoothing out both physical tension and anxiety. The founders, both aestheticians, also say they help clear skin by reducing inflammation and boosting cell turnover.

Suppose you paid the required six dollars per sticker because you really want clear skin. But then your friends started making fun of you for being a sucker. In defending yourself, you'd want to claim that Body Vibes really do have an effect on the clarity of your skin. But what, exactly, would you mean by that claim?

Here's a way of thinking about this. Imagine an alternative world where, at the exact moment you went to stick on your Body Vibes stickers, unbeknownst to you, one of your friends replaced them with identical-looking stickers that cost ten cents instead of six dollars, but which hadn't been "pre-programmed to an ideal frequency." If your skin clarity would be worse in that alternative world, then we would say that Body Vibes have a positive effect on your skin clarity. If your skin clarity would be the same in that alternative world, we'd have to conclude that Body Vibes don't have the claimed effect on skin clarity And if your skin clarity would actually be better in that alternative world, we'd conclude Body Vibes have a negative effect.

We can extend this thought experiment. There's nothing particularly special about the real world. Once we're already thinking about one alternative world, we might as well think about two. For instance, we could think about the effect of ten-cent stickers compared to magical crystals, even if you've never tried either of those approaches to skin care. We just have to compare two make-believe worlds: one where your friends secretly stuck stickers on your upper left shoulder near your heart, and another where

they snuck crystals into your pockets. These kinds of comparisons are called *counterfactual* thought experiments because at least one of the worlds we are comparing isn't the real, factual world—it's in our imaginations. The comparison of outcomes in such a thought experiment is a *counterfactual comparison.*

We can now make sense of the phrase *would result* in our definition of a causal effect. It refers to a counterfactual comparison between the outcome in the actual world and the outcome in a counterfactual world that is identical to the actual world up until the point where the feature of the world claimed to have a causal effect is changed.

This idea of counterfactuals is philosophically subtle. So, to help us make sure we are thinking clearly, we are going to introduce a mathematical framework for representing counterfactuals called *potential outcomes.* Using the potential outcomes framework requires some notation, but it isn't too complicated. And once you master the notation, you will have a much deeper understanding of what causality really is. So let's give it a shot.

Potential Outcomes and Counterfactuals

We are interested in the effect of some *treatment* (say, Body Vibes) on some outcome (say, skin health). Let's call the treatment T. It is a binary variable, taking a value of 0 or 1. If $T = 1$ for some person, that means the person received the Body Vibes treatment. If $T = 0$ for some person, that means the person didn't receive the Body Vibes treatment. We sometimes say that a unit (here, a person) with $T = 1$ is *treated* and a unit with $T = 0$ is *untreated,* although it's often arbitrary what we call treated and what we call untreated (e.g., we could just as easily talk about the effect of *not* wearing Body Vibes).

Similarly, let's refer to the outcome we are interested in as Y. In our example, Y describes a person's skin health. In a metaphysical sense, there is some level of skin health that each individual would have had if they'd used Body Vibes and some level of skin health they would have had if they hadn't used Body Vibes. These are that person's *potential outcomes.* However, at any given moment, we only ever get to observe one of these—each person is either using or not using Body Vibes. Nonetheless, thinking about both potential outcomes helps us to think clearly about counterfactuals:

$$Y_{1i} = \text{outcome for unit } i \text{ if } T = 1$$

$$Y_{0i} = \text{outcome for unit } i \text{ if } T = 0$$

The effect of wearing Body Vibes on person i's skin health is just the difference in i's skin health with and without Body Vibes. In our potential outcomes notation, it is

$$\text{Effect of Body Vibes on } i\text{'s Skin Health} = Y_{1i} - Y_{0i}.$$

Table 3.1 makes this more concrete. We observe ten individuals. For each individual, we observe whether they received Body Vibes and whether their skin is clear. If person i received Body Vibes, their treatment status is $T_i = 1$; if they did not, their treatment status is $T_i = 0$. And if person i had treatment status T, we write their outcome as $Y_{Ti} = 1$ if their skin is clear and $Y_{Ti} = 0$ if their skin is not clear.

The actual outcome for each individual is bold in the table. Individuals 1–5 received Body Vibes, so their actual outcome is Y_{1i}. The table also tells us what these individuals' outcomes would have been if they hadn't received Body Vibes, Y_{0i}. However, in

Table 3.1. Potential outcomes for skin health with and without Body Vibes. For each individual, the actual outcome that we can observe is in bold type. The counterfactual outcome that we do not observe is in regular type.

		Skin Health with Body Vibes Y_{1i}	Skin Health without Body Vibes Y_{0i}	Treatment Effect for Individual i $Y_{1i} - Y_{0i}$
Receive Body Vibes	**Individual 1**	**1**	1	0
	Individual 2	**0**	0	0
	Individual 3	**0**	0	0
	Individual 4	**1**	1	0
	Individual 5	**1**	1	0
Don't Receive Body Vibes	**Individual 6**	0	**0**	0
	Individual 7	0	**0**	0
	Individual 8	1	**1**	0
	Individual 9	1	**1**	0
	Individual 10	0	**0**	0

the actual world, no one can observe these counterfactual outcomes, since they don't actually occur. Individuals 6–10 do not receive Body Vibes. So their actual outcome is Y_{0i}. Again, although the table tells us what their outcomes would have been if they'd received Body Vibes, Y_{1i}, these counterfactual outcomes are not observed in the actual world.

Because the table tells us the potential outcomes in the actual and counterfactual worlds, we can find the treatment effect of Body Vibes for each individual by calculating $Y_{1i} - Y_{0i}$. Doing so reveals that Body Vibes don't actually have any effect on the skin health of any individual. Individuals 1, 4, 5, 8, and 9 all have clear skin. But for all of these individuals, that would be true whether or not they received Body Vibes. Individuals 2, 3, 6, 7, and 10 all have unclear skin. Again, however, this would be true with or without Body Vibes. Importantly, as we will come back to later, this absence of a causal effect can't actually be observed in the world because we only observe the actual outcome for each individual, not the potential outcome in the counterfactual world where they had a different treatment status.

We say that causality is about counterfactual comparisons because we can only observe, at most, one of the two quantities, Y_{1i} or Y_{0i}, for any individual at any particular point in time. This means that we can't directly measure the effect of wearing Body Vibes on an individual's skin health. We suspect this fact is key to their business model.

What Is Causation Good For?

Knowledge of causation is necessary for understanding the consequences of an action that changes some feature of the world. In particular, to weigh the costs and benefits of a decision, you need to know how your action will affect the outcomes you care about.

For instance, you can't possibly know if it is a good idea to spend money on a drug to treat heart disease without knowing about a causal relationship—whether the drug reduces the risk of heart disease. The same goes for many decisions. When you are deciding whether or not to intervene in the world in some way—with a policy, an exercise plan, a parenting strategy, a new kind of online learning, or what have you—you want to know how the intervention *affects* the outcomes you care about.

While the examples we've discussed are easily understood in terms of counterfactual comparisons, sometimes thinking in terms of counterfactuals can seem vexing or confusing. In the next sections, we explore some of these issues.

The Fundamental Problem of Causal Inference

In our discussion of table 3.1 we nodded toward an important issue—causal effects as we've defined them can never, ever be directly observed. Everyone either receives Body Vibes or doesn't receive Body Vibes. So you only observe one potential outcome for each person. But the causal effect is the difference in a person's potential outcomes. This inherent unobservability of causal effects is called the *fundamental problem of causal inference*. Let's see exactly why we can't observe causal effects and what that implies for our ability to learn about causality.

The effect of going to college on your income is the difference in your income in a world in which you go to college versus a world in which you are the same up until the college decision but you don't go to college. At least one of those worlds is counterfactual. You can't both go to college and not go to college. That is, you have two potential outcomes—Y_{college} and $Y_{\text{no college}}$. But you have only one *actual* outcome: either you went to college or you didn't. Given this, we can never observe the effect of going to college on your income since we only observe your income in the actual world, not the counterfactual world.

The fundamental problem of causal inference, then, is that, at any given time, we only observe any given unit of analysis (e.g., a person, basketball team, or country) in one state of affairs. So we can't observe the effect on that unit of being in that state of affairs versus some other state of affairs, because all the other states of affairs are counterfactual. We can't know $Y_{\text{college}} - Y_{\text{no college}}$ for you, because we only observe one of the two values. We saw this fact earlier, in table 3.1, where we noticed that we could only observe the actual outcome for each individual; the other potential outcome was counterfactual.

So how do we make progress on answering causal questions if effects are fundamentally unobservable? Fortunately, there are lots of situations where we don't necessarily need to know the effect for every individual unit of analysis. Instead, we want to know the average effect across lots of individuals.

Suppose, for instance, that the Food and Drug Administration (FDA) is deciding whether to approve a new drug. To learn about the health effects of the drug, scientists conduct a randomized trial, assigning some people to take the drug (the treated group) and other people to take a placebo (the untreated group). Because of the fundamental problem of causal inference, the scientists can't observe the effect of taking the drug on any individual. Each person is either taking the drug or not. But by comparing the average health outcomes for people in the untreated group to the average health outcomes for people in the treated group, they can assess the average effect of the drug. (We'll talk a lot more about how this works in parts 2 and 3.) Doing so allows the scientists

to answer what turns out to be the key causal question for the FDA's decision: If we approve the new drug, how will health change in the population on average?

Drug approval is one setting in which knowledge about average effects is sufficient to inform the key decisions. But there are some settings where this is not the case and the fundamental problem of causal inference constitutes a real challenge. For instance, assessing legal liability involves what's called the *but-for* test. The test requires answering questions like "Would a harm to Anthony not have happened but for Ethan's actions?" The fundamental problem of causal inference says we can never know for sure, since the world in which Ethan did not take his action is counterfactual, so we don't know what happens to Anthony in that world. Instead, what we've just said, and will cover in much more detail in the rest of the book, is that there are methods for answering a slightly different question like "On average, when people take actions of the sort Ethan took, does it tend to cause harm to other people?" A convincing answer to that latter question may or may not be compelling in a court that wants to answer the former.

Part of clear thinking about causal relationships involves admitting that sometimes we cannot answer certain questions with complete confidence, even when those questions are very important.

Conceptual Issues

Causality is a deep and difficult topic. The counterfactual definition of causality doesn't provide all the answers. But it can help us think more clearly about some thorny conceptual issues. Let's talk through a few of these.

What Is the Cause?

One frustration people sometimes feel with regard to the counterfactual approach is that some of the causal questions that we are accustomed to asking appear incoherent within the counterfactual framework. Think of questions like the following: Why did housing prices tank during the latest financial crisis? Why did the Chicago Blackhawks win the Stanley Cup? What caused World War I? Questions of causal attribution like these are common. But when causation is defined in terms of counterfactual comparisons, they don't make a ton of sense.

Let's think about World War I. A common claim is that World War I was caused by the assassination in 1914 of Archduke Ferdinand, the heir to the throne of Austria-Hungary. The assassins were part of a movement that wanted Serbia to take control over the southern Balkans, including Bosnia and Herzegovina, which Austria-Hungary had annexed in 1908. The government of Austria-Hungary responded to the assassination with the July Ultimatum, a list of demands so onerous they were certain to be rejected by the Serbian government. When the ultimatum was rejected, Austria-Hungary declared war on Serbia, leading Russia to mobilize its army to defend Serbia. In response, Germany (an ally of Austria-Hungary) declared war on Russia, France (an ally of Russia) declared war on Germany, and the whole mess cascaded into World War I. Thus, the claim goes, the assassination of Archduke Ferdinand caused World War I.

Now, there is a sense in which this claim is perfectly simple to think about in our framework. We can ask, In the counterfactual world in which Ferdinand was not assassinated, would World War I still have occurred? If World War I would not have occurred in that counterfactual world, then it seems right to say that the assassination had an effect on war breaking out. But that is a far cry from saying that the assassination of

the archduke was *the* cause of the war. Surely, there are many factors that, had they been different, would have prevented World War I from being fought. Sure, had Archduke Ferdinand not been assassinated, maybe the war wouldn't have been fought. But also, had Austria-Hungary not annexed Bosnia and Herzegovina, perhaps Ferdinand would have never been assassinated and the war would have never been fought, so the annexation was just as much a cause as the assassination. Similarly, had the Serbian government complied with the July Ultimatum, perhaps the war would have been avoided, so the noncompliance with the ultimatum was also a cause. And to further illustrate how many such causes there are, had some fish-like creature in the Paleozoic Era swam left instead of right, perhaps the human race as we know it would not exist, and again, World War I would have never been fought. Or, to take an example with some historical gravitas, the seventeenth-century French mathematician Blaise Pascal, reflecting on Mark Antony's attraction to a long proboscis, quipped, "Cleopatra's nose, had it been shorter, the whole face of the world would have been changed."[1] This led James Fearon, in an essay on counterfactual reasoning, to ask, "Does this imply that the gene controlling the length of Cleopatra's nose was a cause of World War I?" As you can see, then, the problem isn't that it is false that the assassination of Archduke Ferdinand caused World War I. Rather, since so many factors appear to have caused World War I, talk of one single cause seems pointless and misguided.

Once we start thinking about counterfactuals, it becomes pretty clear that things have lots of causes. That makes it hard to answer "What is *the* cause" questions. Instead, it pushes us to ask "Was this a cause" or "Did this have an effect" questions. This is perhaps disappointing.

One thought you might have, in response, is that surely some causes of a phenomenon are more important or more proximate than others. If that is true, perhaps we can still talk about the *important* or the *proximate* causes of World War I. How might we do this?

An approach that some philosophers advocate goes something like this. Imagine all the counterfactual worlds in which World War I did not occur. Some of these counterfactual worlds are very different from the actual world—for instance, World War I probably doesn't occur in many counterfactual worlds in which there is no gravity. Others are quite similar to the actual world—perhaps World War I doesn't occur in a world identical to ours through June 27, 1914, but in which Archduke Ferdinand overslept on June 28. We learn about the proximate causes of World War I by comparing the actual world to the counterfactual world in which World War I did not occur that is most similar to the actual world. This kind of analysis may allow us to give reasonable-sounding answers to "What is *the* cause" questions without abandoning our definition of causation based on counterfactual comparisons. For instance, it seems reasonable to think that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra's nose, the laws of gravity, or the whims of Paleozoic fish.

There is certainly something to this approach. But, that said, it is often hard to assess the importance or proximity of one cause versus another in a principled way. If you know a bit of history, you surely can come up with other causes of World War I that seem equally proximate. For instance, many scholars have argued that early-twentieth-

[1] Antony and Cleopatra's love affair had major repercussions for world history. For instance, historians generally believe that the end of the Roman Republic and the establishment of the Roman Empire were ensured when Antony and Cleopatra were defeated by Octavian (later, Emperor Augustus) at the Battle of Actium. Had this not occurred, who knows how differently the rest of western history might have played out?

century military doctrines favoring offensive over defensive strategies played a role in causing World War I. Is the world in which a slightly different military doctrine was adopted more proximate to our world than the one in which Archduke Ferdinand was not assassinated? For that matter, is the world in which one Paleozoic fish took a different turn really such a large leap? It's hard to say.

To see the problem in a somewhat less lofty and perhaps more familiar setting, consider an NCAA Division III women's basketball game between the Chicago Maroons (where some of our star students are also star athletes) and the Emory Eagles. Suppose the Maroons are trailing the Eagles by one point, and the Maroons have just enough time left to take one final shot. They make it, winning the game by one point (in basketball, field goals are worth at least two points). The next day, the *Chicago Maroon* newspaper will fixate on that last shot, and the reporter might even write that the last shot was *the* reason the Maroons won.[2] But think about this counterfactually for a moment. Dozens of shots throughout the game were pivotal. Plausibly, every shot the Maroons made was pivotal—in a counterfactual world in which they missed that shot and everything else played out as it did in the actual world, they would have lost instead of won. Similarly, every shot the Eagles missed was pivotal—in a counterfactual world in which they made it and everything else played out as it actually did, they would have won instead. So what's so special about that last shot? One possibility is that everyone knew that the final shot would be pivotal when it was taken. But very few other causes meet this criterion, certainly not the assassination of Archduke Ferdinand. So, in our view, there is no obvious reason to think that the last shot was a more important cause of the Maroons' victory than the other shots. Instead, we think this example illustrates a basic, if frustrating, fact of life: individual events can have many equally important and consequential causes.

Another surprising fact about the counterfactual approach is that, at least in principle, it's possible for some event to have no causes at all. Suppose that the authors of this book concoct the perfect crime. We both shoot and kill our sworn enemy at the same time, knowing that either bullet would be fatal on its own. When questioned, Anthony says, "Clearly, I can't be charged with a crime. My actions had no effect whatsoever. Had I not fired my gun, the victim would still have died." And similarly, Ethan retorts, "I could not have possibly caused the victim's death either. Had I not shot my gun, he would have still died." While the justice system might not be impressed by our defense, the counterfactual logic is sound. Some events may be the result of a confluence of factors whereby no single factor could have changed the outcome. This theoretical possibility is yet another reason that it might not make much sense to ask questions like "What caused World War I?" It could well be that, for all the factors we like to talk about, taking away any one of them would in fact not have sufficed to prevent the war.

Causality and Counterexamples

One common skeptical reaction to evidence showing the existence of an average effect is to point to counterexamples. Perhaps you've had an experience like the following at a family gathering. You read a study showing that, on average, flu shots reduce the risk of contracting the flu. You mention this over Thanksgiving dinner, encouraging

[2] We know it's confusing that the basketball players are the Maroons, the newspaper is the *Maroon*, and probably neither sports teams nor newspapers should be named after a color. Our university is typically not known for athletics or branding.

your loved ones to get the vaccine. But your vaccine-skeptic relative says, "I don't know, I got the flu shot last year and I still got the flu." Many people nod and agree, perhaps pointing out that their friend so-and-so also got the flu shot and still got sick.

The intuition behind this kind of objection-by-way-of-counterexample is something like this: "If flu shots really prevent the flu, then no one who got a flu shot would get the flu. Thus, my one counterexample means the vaccine doesn't work."

This argument does not reflect clear thinking. The evidence says that the flu shot caused flu risk to go down, averaging across lots of people, each with their unique biology, level of flu exposure, environment, and so on. It doesn't say that it eliminated flu risk for each and every individual. But to get flu risk to go down on average, the flu shot must have prevented the flu (i.e., had a causal effect) for at least some people. We just don't know exactly which ones experienced the effect.

Let's think about this in our potential outcomes notation. Think of the potential outcomes as whether or not you get the flu. We'll say $Y = 1$ if you stayed healthy and $Y = 0$ if you got the flu. And think of the treatment as whether you got the flu shot, with $T = 1$ meaning you got the shot and $T = 0$ meaning you didn't.

Maybe there are three different kinds of people—call them the *always sick*, the *never sick*, and the *vaccine responders*. The always sick and the never sick have potential outcomes that don't respond to treatment. The always sick get the flu regardless of whether they get the flu shot, and the never sick never get the flu. In our notation,

$$Y_{1,\text{always sick}} = 0 \qquad Y_{0,\text{always sick}} = 0$$

and

$$Y_{1,\text{never sick}} = 1 \qquad Y_{0,\text{never sick}} = 1$$

But the vaccine responders are different; they get the flu if they don't get the shot, and they don't get the flu if they do get the shot:

$$Y_{1,\text{vaccine responder}} = 1 \qquad Y_{0,\text{vaccine responder}} = 0$$

In a population made up of these three groups of people, getting the flu shot reduces the probability you will get the flu. That is, on average, the treatment effect is positive. You don't know which group you are in. There is a chance you are a vaccine responder. So getting a flu shot reduces your probability of getting sick.

Let's see this in an example. Suppose there are 10 individuals. Individuals 1–5 get the flu shot, while individuals 6–10 don't. Individuals 1, 3, 4, 5, and 8 are always-sick types, so they get the flu. Individuals 5, 6, 7, and 10 are never-sick types, so they stay healthy. Individuals 2 and 9 are vaccine responders. Individual 2 gets the flu shot, so she stays healthy. But individual 9 does not get the flu shot, so he gets sick.

Table 3.2 shows potential outcomes and treatment effects. As we can see, not everyone in this population has a positive treatment effect. But the average of the treatment effects across these 10 individuals is $\frac{2}{10}$ because two of the ten are vaccine responders. So, for any individual, not knowing which type of person they are, there is a 20 percent chance that taking the flu shot will prevent them from getting the flu.

Importantly, pointing to one counterexample is neither here nor there with respect to such evidence. Perhaps your unlucky relative was a person, like individual 1, 3, or 4, whose confluence of circumstances were such that the flu shot didn't have an effect (i.e., they were an always sick). That doesn't mean it didn't have an effect for other people.

Table 3.2. Potential outcomes for flu with and without the flu shot. For each individual, the actual outcome that we can observe is in bold type. The counterfactual outcome that we do not observe is in regular type.

		Health with Flu Shot Y_{1i}	Health without Flu Shot Y_{0i}	Treatment Effect for Individual i $Y_{1i} - Y_{0i}$
Flu Shot	**Individual 1** (always sick)	**0**	0	0
	Individual 2 (vaccine responder)	**1**	0	1
	Individual 3 (always sick)	**0**	0	0
	Individual 4 (always sick)	**0**	0	0
	Individual 5 (never sick)	**1**	1	0
No Flu Shot	**Individual 6** (never sick)	1	**1**	0
	Individual 7 (never sick)	1	**1**	0
	Individual 8 (always sick)	0	**0**	0
	Individual 9 (vaccine responder)	1	**0**	1
	Individual 10 (never sick)	1	**1**	0

And it doesn't even mean that the flu shot won't prevent the flu for that same relative next year or that it won't help you. Absent any further information about which group they are in, any individual's best guess is that the flu shot will reduce their chances of contracting the flu since it does so on average. And we haven't even discussed the more complicated issue that outcomes aren't actually binary, so the shot may have a causal effect on the severity of the flu.

Of course, the possibility that effects are different for different people presents another set of important conceptual challenges. We might be able to detect such *heterogeneous treatment effects*, especially if they correspond with observable categories (e.g., men versus women, older versus younger, healthy versus sick). To identify such heterogeneous effects, we could run a separate experiment for each group, which would tell us the average effect for each group rather than for the whole population. But what if effects differ across people for complicated or obscure reasons that might never occur to us? Then, when we go to look at the effect of some intervention, it is very important to keep in mind that we are learning about an average effect. Some people may have effects much larger than the average. Others may have effects much smaller than the average. Indeed, some people may have no effect at all or an effect in the opposite direction from the average. If we don't know the source of this heterogeneity, all we will

be able to say is something about the average, which, as we've discussed, may still be valuable.

Causality and the Law

As we briefly mentioned previously, one place where philosophical questions about causality become of serious practical import is in the law. Administering justice requires assigning blame and assessing liability. If we want to know whether, say, Ethan should be held liable for some harm suffered by Anthony, surely we need to know whether Ethan's actions caused that harm. But, as we've just discussed, talking about causes in this way is conceptually fraught. Many things, from the behavior of a Paleozoic fish to Ethan's alleged negligence, may have had a causal effect on the harm Anthony suffered. Is the fish liable too?

The law is aware of the philosophical conundrum. But it must ultimately come up with some pragmatic resolution that allows judges and lawyers to get on with the business of administering justice. Here's, roughly, where it comes down.

In the Common Law, causality is thought of in terms of two conditions that are closely related to things we've talked about. These are referred to as *cause-in-fact* and *proximate causality*.

Cause-in-fact is essentially counterfactual causality. Whether Ethan's actions are a cause-in-fact of Anthony's suffering is determined by whether Anthony wouldn't have suffered *but for* Ethan's actions.

Of course, as you already know, a counterfactual standard like the but-for test isn't very stringent. World War I wouldn't have happened but for a Paleozoic fish turning the wrong direction. Does that mean we should blame the poor fish for World War I?

The law's answer is no. The fish is off the hook, so to speak. This is where proximity comes in. For there to be liability, the law requires that some cause-in-fact be close enough in the causal chain. This thought is also familiar—for instance, from our argument that the assassination of Archduke Ferdinand is a more proximate cause of World War I than is Cleopatra's nose.

So an assessment of legal causality might go something like this. Suppose you order food delivery and the delivery person drives recklessly, crashing into your neighbor's car. Are you liable for your neighbor's suffering? It is plausible that, but for your decision to order delivery, the delivery person wouldn't have been in the area and your neighbor's car wouldn't have been hit. So your actions are probably a cause-in-fact of your neighbor's suffering. But there are many steps in the causal chain between your actions and the car crash, all of which are out of your hands. So the law would not find you liable for the damage to your neighbor's car.

Of course, as we've discussed, knowing exactly how to apply the conditions of cause-in-fact and proximate causality is tricky. To apply the but-for test, we have to know what the right counterfactual world is. And defining how close is close enough for a proximity test is a fraught problem, full of judgment calls. All of which is to say that these questions about causality are vexing and of great practical importance.

Can Causality Run Backward in Time?

One common intuition is that causality must run forward in time. That is, an event that happens now can have an effect on events that happen in the future. But surely, the thought goes, events that happen in the future can't affect events in the past. Indeed,

one common strategy for trying to establish a causal relationship is to show that the supposed cause typically occurs prior to the supposed effect.

Let's check this intuition by thinking about birthday cards. Here's a correlation that we hope is true in the world: the number of birthday cards that get mailed to you in a given week is strongly correlated with it being within a week of your birthday. That is, many more birthday cards are mailed to you in the week before your birthday than in any other week of the year.

Now, although correlation need not imply causation, we suspect that there is a causal relationship here but not the one that's implied by thinking of causal relationships as running forward in time. Receiving birthday cards does not cause your birthday to occur. In a counterfactual world in which those cards were sent at a different time, or even in a counterfactual world in which greeting cards cease to exist, your birthday will still occur on the date you were born. Instead, you might say the causal relationship runs backward in time. Your birthday exerts an effect on the sending of birthday cards. In the counterfactual world in which your birthday occurs in a different month, you will be sent fewer birthday cards in the week preceding your birthday in this world. Thus, on our counterfactual definition, your birthday exerts a causal effect on birthday cards. Causality appears to run backward in time.

There are objections to this line of argument. For instance, one might argue that it isn't your future birthday, but anticipation of that birthday, that exerts a causal effect on the sending of birthday cards. If we changed people's beliefs about whether your birthday is coming up, we'd change their card-sending behavior. But if we changed your actual birthday, without a change in their beliefs, the cards would still be sent. On this argument, causality is operating forward in time, in the intuitive way.

Even that need not be the end of the argument. After all, where did the anticipation of your birthday come from? It presumably came from the fact of your actual birthday. If we changed the fact of your actual birthday in the future, we'd change people's anticipation of your birthday now (which would, in turn, change their card-sending behavior). Perhaps we are back to causality running backward in time. Or perhaps not. Is it really the changing of your birthday in the future that affects people's anticipation today? Or is it telling them about the change in your future birthday, in which case we are right back to causality running forward in time.

As you can no doubt tell by this point, we aren't going to solve this issue here. But we do want you to see two things clearly. First, evidence that one thing occurred before another is not, on its own, convincing evidence that the one caused the other. Second, whether or not you think causality can or cannot run backward in time, we can always define the causal effects in terms of a counterfactual.

Does Causality Require a Physical Connection?

Another intuition many people share is that causation necessarily has to do with physical connection—a view that we'll refer to as *physicalism*. One billiard ball affects another by bumping into it. Maybe such physical connections always underlie causal relationships.

While, of course, there are many examples of causal effects that occur through physical connection, there are good arguments to suggest such physical connection is not required. Think of a person who is deterred from robbing a bank by worry about imprisonment. Such a person's behavior is affected by the existence of the police, the courts, the penal code, and the prison system. The criminal justice system affects whether this person commits a crime, even though there is no physical connection between them.

Indeed, think of our previous discussion of the effect of birthdays on the sending of birthday cards. Birthdays aren't a physical thing in the world at all. It is hard to see what it would even mean for the causal relationship between birthdays and the sending of birthdays cards to occur through physical connection.

A defender of physicalism might say that with enough creativity, we can describe the effect of the criminal justice system on crime in purely physical terms. Perhaps the past arrest and conviction of people who committed crimes led reporters to write about this activity in newspapers, which led the person in question to read about these arrests in the newspaper, which, through a complicated sequence of light hitting the person's eyeballs, led to lots of chemical and electrical connections in that person's brain, which deterred them from committing a crime. You could do a similar exercise for birthdays and birthday cards.

Again, we aren't going to provide a definitive answer. There may be reasonable arguments on both sides of the physicalism debate. The important point is that we can think about counterfactually defined causal relationships that do not depend on anything like the simple, commonsense kind of physical connections suggested by the billiard ball example.

Causation Need Not Imply Correlation

We've agreed that correlation need not imply causation. But, perhaps more surprisingly, causation also need not imply correlation and certainly not correlation in the expected direction. There are many situations in which some feature of the world has (say) a *negative* effect on some other feature of the world, but those two features of the world are *positively* correlated (or vice versa).

You'd probably find a strong, positive correlation between the number of firefighters who have recently visited a house and the amount of fire damage to that house. But if we had to guess, we'd suspect that firefighters, on average, reduce fire damage. In other words, if fewer firefighters had visited, we suspect there would be even more fire damage.

So why is the correlation positive? Firefighters tend to visit houses that are on fire. So, although firefighters reduce fire damage to some degree, the houses that have been visited by firefighters tend to have more fire damage. Hence, not only should one not conclude from a correlation that there must be a causal relationship, but one also should not assume that just because a causal relationship exists, the correlations found in the world will correspond to those causal relationships in some straightforward way.

Wrapping Up

Understanding whether a causal relationship exists is one of the fundamental goals of quantitative analysis. But, if we are going to do that, we need to think clearly about what causality means.

We believe that the best way to conceptualize causality is through a thought experiment involving counterfactuals. A treatment has a causal effect on an outcome if the outcome would have been different had the treatment been different. Of course, in the actual world, the treatment was what it was. We can't observe the counterfactual world in which the treatment was different in order to figure out if the outcome would have been different. This is the fundamental problem of causal inference.

The fact that causal effects are unobservable doesn't mean data analysis cannot help us learn about them. In particular, we can learn about the average effect in some population, even though we can't observe any of the individual effects directly.

Doing so involves making careful use of quantitative knowledge about things like correlations. In part 2 we turn to a more detailed discussion of how we establish and quantify correlations. This will set us up to be able to think clearly in part 3 about estimating causal effects.

Key Terms

- **Causal effect:** Informally, the change in some feature of the world that would result from a change to some other feature of the world. Formally, the difference in the potential outcomes for some unit under two different treatment statuses.
- **Body Vibes:** Stickers that a company called Goop claims cause clear skin. The authors of this book do not endorse Body Vibes, mainly because we will be releasing our own competitor: Brain Vibes. One sticker applied to the temple causes clear thinking.
- **Counterfactual comparison:** A comparison of things in two different worlds or states of affairs, at least one of which does not actually exist.
- **Treatment:** Terminology we use to describe any intervention in the world. We usually use this terminology when we are thinking about the causal effect of the treatment, so we want to know what happens with and without the treatment. Importantly, although it sounds like medical terminology, *treatment* as we use it can refer to *anything* that happens in the world that might have an effect on something else.
- **Potential outcomes framework:** A mathematical framework for representing counterfactuals.
- **Potential outcome:** The potential outcome for some unit under some treatment status is the outcome that unit would experience under that (possibly counterfactual) treatment status.
- **Fundamental problem of causal inference:** This refers to the fact that, since we only observe any given unit in one treatment status at any one time, we can never directly observe the causal effect of a treatment.
- **Heterogeneous treatment effects:** When the effect of a treatment is not the same for every unit of observation (as in the case of flu shots and virtually every other interesting example of a causal relationship), we say that the treatment effects are heterogeneous. Sometimes we're still interested in the average effect even though we know the treatment effects are heterogeneous, and sometimes we want to explicitly study the nature of the heterogeneity. (In contrast, when discussing the unlikely possibility that treatment effects are the same for every unit, we would refer to *homogeneous* treatment effects.)

Exercises

3.1 Sarah says that she is hungry. John hands her a piece of pizza. Sarah eats the pizza and then declares that she is no longer hungry.

(a) The fundamental problem of causal inference seems to say that you can't know that Sarah eating the pizza had a causal effect on her no longer being hungry. Is that right? Explain.

(b) Do you think you nonetheless have good reasons to believe that eating the pizza had an effect on Sarah no longer being hungry? Explain why or why not.
(c) Do you have good reasons for believing that John handing Sarah the pizza had a causal effect on her no longer being hungry? In your assessment, are the reasons to believe John's actions had a causal effect better or worse than the reasons to believe Sarah eating the piece of pizza had a causal effect?

3.2 A government is considering making alcohol consumption illegal as part of a public health campaign. Let's think of making alcohol illegal as the treatment T. Write $T = 1$ if the government makes alcohol illegal and $T = 0$ if the government leaves alcohol legal.

We will think of a binary outcome for each person: either they drink alcohol or they do not. If person i drinks at treatment status T, we write her potential outcome as $Y_{Ti} = 1$, and if she doesn't drink, we write it as $Y_{Ti} = 0$.

Suppose the society is made up of three groups: the always drinkers, the legal drinkers, and the never drinkers. The always drinkers will drink whether or not alcohol is legal. The legal drinkers will drink if and only if alcohol is legal. The never drinkers won't drink whether or not alcohol is legal.

(a) Write down, in potential outcomes notation and as a number (0 or 1), each of the two potential outcomes for each of the three groups.
(b) Write down, in both potential outcomes notation and as a number (0 or 1), the causal effect of making alcohol illegal on drinking for each of the three groups.
(c) Is there an effect, on average, of banning alcohol in this society?
(d) Suppose you are out to lunch with some friends and one of them says, "My uncle lives in a place where they banned alcohol and all of his friends kept drinking, so I don't think the ban does anything." Explain, in terms of our example, why this isn't a convincing argument.

3.3 The Republican National Committee (RNC) has hired three consultants and asked them to figure out the cause of their loss in the 2020 presidential election. The first consultant says that they didn't do enough television advertising. The second consultant reports that they should have encouraged more of their supporters to vote rather than criticizing voting by mail. The third consultant concludes that Donald Trump should have done a better job responding to the COVID-19 pandemic and should have shown more compassion on the campaign trail. Confused by the apparently conflicting information, the RNC hires you, a quantitative analyst, to adjudicate between these three possibilities. What would you tell them? How would you proceed?

3.4 In the 2016 U.S. Open golf tournament, Dustin Johnson was leading the tournament in the final round, and his ball was resting on the fifth green. While preparing for his upcoming putt, he tapped his putter on the ground next to the ball and the ball moved. The rules at the time stated that if we were highly certain that a player caused his ball to move, even if it was inadvertent, he or she should incur a penalty. Because you're an expert on causation, the rules

officials call you in to evaluate the situation. The officials make the following arguments. Please provide your expert response to each one.

(a) Johnson couldn't have possibly caused the ball to move, because he (and his putter) never touched it.
(b) Johnson shouldn't receive a penalty because the true cause of the ball moving was the greenskeeper. Had the greenskeeper not cut and rolled the greens so much that morning, the ball wouldn't have moved.
(c) An empirically minded official went out to the same green, placed a ball down, tapped his putter on the ground next to the ball, and it didn't move. Therefore, Johnson's actions couldn't have caused the ball to move.
(d) One official was watching the incident up close and says he's virtually certain that if Johnson had not tapped his putter next to the ball, it wouldn't have moved. Therefore, he caused it to move and should receive a penalty.

Readings and References

You can read about Body Vibes on the Goop website. We last accessed it on June 15, 2020. http://goop.com/wearable-stickers-that-promote-healing-really/

The quote from Blaise Pascal on Cleopatra's nose is from his seventeenth-century collection entitled *Pensées*.

The essay about counterfactual reasoning discussing the gene controlling the length of Cleopatra's nose is

James D. Fearon. 2011. "Counterfactuals and Hypothesis Testing in Political Science." *World Politics* 43(2):169–195.

If you'd like to read more about the counterfactual definition of causality, potential outcomes, and surrounding discussions and debates, have a look at these:

David Lewis. 1973. "Causation." *Journal of Philosophy* 70:556–67.

Paul W. Holland. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60.

Stephen Mumford and Rani Lill Anjum. 2014. *Causality: A Very Short Introduction.* Oxford University Press.

There is also a nice entry by Peter Menzies and Helen Beebee in the *Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/entries/causation-counterfactual/.

PART II

Does a Relationship Exist?

CHAPTER 4

Correlation Requires Variation

What You'll Learn

- You can't learn about a correlation without variation in both variables of interest.
- In many realms of life—from education to medicine to rocket science—people fall into the trap of trying to make claims about correlations without such variation.
- A particularly common way people fall into this mistake is by *selecting on the dependent variable*, examining only instances when some phenomenon occurred rather than comparing cases where it occurred to cases where it did not.
- Many institutional procedures push us to select on the dependent variable without noticing it.

Introduction

In chapter 2 we discussed the idea that the correlation between two features of the world is the extent to which they tend to occur together. We opened our discussion of correlation by thinking about whether oil production and autocracy are correlated. To figure this out we looked at the country-level data represented in table 4.1.

To determine whether there is a correlation between oil production and autocracy we compared the percentage of major oil producers that are autocracies to the percentage of countries that aren't major oil producers that are autocracies. To make this comparison, we needed four pieces of information: the number of autocracies that are major oil producers, the number of democracies that are major oil producers, the number of autocracies that are not major oil producers, and the number of democracies that are not major oil producers. Had we been lacking any of these pieces of information, we would not have been able to figure out whether oil production and autocracy are correlated.

To see why, suppose we didn't know the number of democracies that are major oil producers. (Of course, we'd also have to not know the total number of countries, so we couldn't just back out the 9 by subtracting the number of countries in the other three categories from the total number of countries.) We still know that about 20 percent ($\frac{29}{147}$) of countries that aren't major oil producers are autocracies. But now we can't figure out

Table 4.1. Oil production and type of government.

	Not Major Oil Producer	Major Oil Producer	Total
Democracy	118	9	127
Autocracy	29	11	40
Total	147	20	167

what proportion of the major oil producers are autocracies. It could be anything. If the number of democracies that are major oil producers turned out to be (say) 11, then 50 percent ($\frac{11}{22}$) of major oil producers would be autocracies and there would be a positive correlation. If the number of democracies that are major oil producers turned out to be (say) 99, then only 10 percent ($\frac{11}{110}$) of major oil producers would be autocracies, so there would be a negative correlation. If the number of democracies that are major oil producers turned out to be 44, then 20 percent ($\frac{11}{55}$) of major oil producers would be autocracies—the same as for countries that are not major oil producers—and there would be no correlation at all. So, just as we saw in our discussion of scandals and congressional representatives in chapter 2, we need to observe all four pieces of information to figure out the correlation.

This is what we mean when we say that correlation requires variation: If you want to figure out whether two variables are correlated, you have to observe variation in both of them. You must observe the number of countries that are and are not major oil producers. And you must observe the number of autocracies and democracies in each group. Just observing variation in one or the other variable is not enough. In chapter 2, when we asked which of five factual statements described a correlation, the problem with the three statements that did not was a lack of variation in one of the variables.

While it may seem obvious, on the basis of our simple binary example, that correlation requires variation, in our experience, it is anything but. Indeed, failing to look for variation in one or another variable while trying to establish a correlation is an exceptionally common mistake.

In this chapter, we explore this mistake and try to unpack why it is so common. Broadly, we think there are two closely related reasons that people so frequently try to establish a correlation without variation. The first reason is called *selecting on the dependent variable.* The second reason is that the world is often organized in ways that push us to make this mistake.

This chapter, more than most in the book, is built around examples. We do this for a reason. We've found that, once we explain that correlation requires variation, people tend to nod their head in agreement, appearing to understand. Indeed, because the point seems obvious when put in plain English, many people are skeptical that this could be such a big problem. And yet, they themselves go right back to making the same mistake. We hope that by showing you lots of examples of very smart people making this mistake in high-stakes environments, we will convince you that this is a real problem and that avoiding this error requires clear thinking, genuine effort, and concentration.

Selecting on the Dependent Variable

If you want to forecast or explain some phenomenon, it is a natural impulse to start by examining previous instances of that phenomenon occurring. This is called *selecting on the dependent variable.* But if you look only at instances when the phenomenon

occurred, you are trying to assess a correlation without variation, since you have no variation in whether or not the phenomenon occurred. This is like looking for correlates of autocracy without examining any democracies. It won't work.

The phrase *dependent variable* refers to the variable representing the phenomenon you are trying to forecast or explain. This mistake is referred to as *selecting on the dependent variable* because you are selecting which cases to look at based on the value of the dependent variable (e.g., only looking at autocracies) rather than looking at variation in the dependent variable (e.g., comparing autocracies and democracies).

Consider a few examples. Following the financial crisis of 2008, both scholars and journalists who wanted to understand how to predict future financial crises invested enormous time and energy examining the historic record to look for patterns in previous crises. Malcolm Gladwell, in his book *Outliers*, tries to understand the correlates of personal success by recounting the lives of highly accomplished people, looking for similarities. Congress, considering a change to American counterinsurgency strategy in Afghanistan, heard testimony on the correlates of suicide terrorism from an academic expert who had done an exhaustive study of all suicide terrorist campaigns since 1980, looking for shared characteristics.

As natural as it seems to look for commonalities in past instances of events you want to forecast, it really is a mistake. Correlation requires variation. Each of the studies just described would have been far more informative if they'd had variation in the dependent variable.

The claim that we can't learn about the correlates of financial crises or suicide terrorism by looking for commonalities among historic cases of similar events may seem counterintuitive. But, since we know that correlation requires variation, the mistake is actually quite simple to grasp. Put in the terms of our earlier example, each of these examples is analogous to looking for correlates of oil production without any data on non-oil-producing countries!

To see the key conceptual flaw in all of these arguments in another way, let's start by considering the central claim in Gladwell's *Outliers*, the so-called *10,000-hour rule*.

The 10,000-Hour Rule

Gladwell's idea is that it takes about 10,000 hours of serious practice to master any difficult skill. Talent might matter too, but first and foremost, if you are looking for a great achiever, look for someone who put in that 10,000 hours of practice.

Now, of course, Gladwell isn't just interested in forecasting great success. He thinks the 10,000-hour rule might be causal. If true, this would have far-reaching consequences. Given enough practice, perhaps any of us could achieve almost anything.

But talk of causality is premature. Before we can think about causality, we need to figure out whether Gladwell's evidence is even compelling for the claim of a correlation between 10,000 hours of practice and great success. So let's start there.

Gladwell asks, "Is the ten-thousand-hour rule a general rule of success?" The answer, he concludes, is yes. The evidence? "If we scratch below the surface of every great achiever" we see the same pattern (p. 47). "Virtually every success story... involves someone or some group working harder than their peers" (p. 239). In case after case, from Bill Gates to the Beatles, Gladwell shows that great achievers put in their 10,000 hours—overwhelming evidence, he concludes, that practice predicts success.

Let's try to think a little more clearly about Gladwell's evidence. What has Gladwell shown us? Of course, he hasn't actually looked at every great achiever. But he's shown us evidence that lots of great achievers practice at least 10,000 hours. The big problem

Table 4.2. Great achievers practice more than 10,000 hours.

	Great Achiever	Not Great Achiever	Total
10,000 Hours of Practice	Many	?	
Not 10,000 Hours of Practice	Very few	?	
Total			

is that he's told us nothing about all the people who aren't great achievers. A table of evidence for *Outliers* would look something like table 4.2.

Even granting that Gladwell is correct that most great achievers put in 10,000 hours of practice, this doesn't tell us whether 10,000 hours of practice is correlated with great success. Correlation requires variation. Because he has selected on the dependent variable, Gladwell's data lack variation in achievement. If you want to know whether putting in 10,000 hours of practice correlates with success, it is not enough to observe that most great achievers put in 10,000 hours of practice. We need to know about the non-achievers' practice habits as well.

Of course, Gladwell's analysis does provide some information that we didn't previously have. Momentarily, let's suppose that Gladwell didn't cherry pick his stories in order to fit his narrative (although, of course he did: he's a storyteller, not a scientist). In this case, we've learned that most highly successful people put in 10,000 hours of practice before achieving great success.

Although this is not enough information to measure a correlation, Gladwell and his defenders might argue that we already have a rough sense that most members of the general public who are not great achievers have not put in 10,000 hours of practice. In that case, maybe Gladwell's analysis significantly shifts our beliefs about the correlation between practice and great success, even if he didn't explicitly measure the correlation. In these cases where we already have a good sense of the prevalence of something in the general population, perhaps it's useful to show that the prevalence is different for a certain group of interest.

Maybe. But we're still skeptical that Gladwell's analysis teaches us much. That's because most people probably *have* devoted at least 10,000 hours of practice to *something*. Anthony has spent 10,000 hours on the golf course, and he's no Tiger Woods. Ethan has spent 10,000 hours playing guitar, and he's no Jimi Hendrix. If you've worked at something full time for five years but you're not the most successful person in your field, then you're one of the many, many people in the top-right cell of table 4.2 that Gladwell never considered.

We should also remember that Gladwell is a gifted storyteller. In the extremely unlikely scenario in which Anthony wins the Masters, Gladwell might write an inspiring and convincing story about how, despite being a full-time college professor, Anthony's many years of practice, failure, and more practice allowed him to pull off the greatest Cinderella story in sports history (just let us dream for a moment). But far more likely, Anthony will happily continue to be one of millions, if not billions, of people who love something, work hard at it, but never achieve immense success and who are never considered in Gladwell's analysis.

To test your understanding, let's see the problem with claims like Gladwell's in another setting. We are going to repeat his exact argument, but in a fictional example that we hope makes the problem even clearer.

Table 4.3. What sick people drank (made-up data).

	Sick	Not Sick	Total
Drank Beverage	500		
Didn't Drink Beverage	0		
Total	500		

Table 4.4. What sick and healthy people drank (made-up data).

	Sick	Not Sick	Total
Drank Beverage	500	9,500	10,000
Didn't Drink Beverage	0	0	0
Total	500	9,500	10,000

Suppose a town of 10,000 people experiences a surprising spate of illness. In the course of a month, 500 people are taken ill with the same symptoms. Local health officials want to determine the cause of the illness. They take case histories of the 500 sick people, looking for commonalities. In the course of this investigation, they find that all 500 people consumed the same beverage, from the same source, the day before they were hospitalized.

Table 4.3 shows data corresponding to our fictionalized story.

The facts about the beverage and the illness correspond exactly to the facts about practice and success from *Outliers*. Everyone who gets sick (succeeds) drank the same beverage (put in 10,000 hours). Surely, then, drinking that beverage (practicing 10,000 hours) is an important predictor of illness (great success). If we want to know who else is likely to get sick, we should survey the town and find out who else drank the same beverage. Right?

Suppose we tell you that the beverage in question is tap water. The claim that the "pattern" of illness suggests a correlation between the beverage and the disease now seems questionable. Why? Because many people consume tap water every day. Indeed, in our fictional town, all 500 people who got sick consumed tap water, but so too did the 9,500 who didn't get sick. As table 4.4 makes clear, there is in fact no correlation between the beverage and getting sick: 100 percent of sick people and 100 percent of healthy people drank the beverage.

The 10,000-hour rule is similarly unsubstantiated by data of the sort presented by Gladwell. Yes, lots of successful people practice very hard. So too do lots of less successful people. Think of all the bands that practiced countless hours, played countless gigs, and did not become the Beatles.

Corrupting the Youth

American kids who liked rock music in the 1980s (ask your parents) may remember the Parents Music Resource Center (PMRC). The PMRC was a lobbying group whose members opposed what they perceived to be the increasingly inappropriate content of rock music. Most famous among the founders of the PMRC was Tipper Gore, wife

of then Senator and later Vice President Al Gore, who started the group after being shocked by the lyrics of a Prince song.

The PMRC claimed that explicit lyrics were corrupting the youth, causing suicide, sexual violence, and even murder. They denounced "porn rock"—a category that included Bruce Springsteen because the song "I'm on Fire" contained a sexual innuendo—and demanded warning labels be placed on albums. In 1985, the Senate Commerce, Science, and Transportation Committee held hearings. Musicians from across the musical spectrum, from the country singer John Denver to Twisted Sister's Dee Snider testified against the PMRC's position. But the PMRC prevailed.

Let's consider a bit of the argument. Here is the testimony of Jeff Ling, a PMRC consultant:

> Many albums today include songs that encourage suicide, violent revenge, sexual violence, and violence just for violence's sake. . . This is Steve Boucher. Steve died while listening to AC/DC's "Shoot to Thrill." Steve fired his father's gun into his mouth. . . A few days ago I was speaking in San Antonio. The day before I arrived, they buried a young high school student. This young man had taken his tape deck to the football field. He hung himself while listening to AC/DC's "Shoot to Thrill." Suicide has become epidemic in our country among teenagers. Some 6,000 will take their lives this year. Many of these young people find encouragement from some rock stars who present death as a positive, almost attractive alternative. . .Of course, AC/DC is no stranger to violent material. . .One of their fans I know you are aware of is the accused Night Stalker.

Ling's argument, which is typical of crusaders against corruption of the youth, amounts to this:

1. Some young people behave regrettably.
2. The youth who behave regrettably all listen to this terrible rock music.
3. The music must be the cause of the regrettable behavior

Of course, talk of causality is again premature. We'll focus on whether such evidence even suggests a correlation.

Thirty years earlier, in 1954, the Senate heard astoundingly similar testimony about that generation's scourge of the youth, comic books. Here is the neurologist and psychiatrist Fredric Wertham testifying before a Senate subcommittee:

> There is a school in a town in New York State where there has been a great deal of stealing. Some time ago some boys attacked another boy and they twisted his arm so viciously that it broke in two places, and, just like in a comic book, the bone came through the skin.
>
> In the same school about 10 days later 7 boys pounced on another boy and pushed his head against the concrete so that the boy was unconscious and had to be taken to the hospital. He had a concussion of the brain.
>
> In this same high school in 1 year 26 girls became pregnant. The score this year, I think, is eight. Maybe it is nine by now.
>
> Now, Mr. Chairman, this is what I call ethical and moral confusion. I don't think that any of these boys or girls individually vary very much. It cannot be explained individually, alone.
>
> Here is a general moral confusion and I think that these girls were seduced mentally long before they were seduced physically, and, of course, all those people

> there are very, very great—not all of them, but most of them, are very great comic book readers, have been and are.

This kind of argument persists in the contemporary environment. We have all heard, and perhaps even made, similar claims about the insidious effects of television or video games or social media. For instance, following the horrific shootings at Columbine High School, the U.S. Department of Education and the Secret Service set up a joint task force to determine what factors would allow school officials to anticipate and prevent school violence. The task force studied all thirty-seven incidents of school violence from 1974 through 2000. While concluding that there is no single profile of a school shooter, they also reported the following (among many other things):

1. "Many attackers felt bullied, persecuted, or injured by others prior to the attack."
2. "Most attackers were known to have had difficulty coping with significant losses or personal failures."
3. "Most attackers engaged in some behavior, prior to the incident, that caused others concern or indicated a need for help."
4. "Over half of the attackers demonstrated some interest in violence, through movies, video games, books, and other media."

A similar commission was convened in 2018. While less focused on specific corrupters of the youth, this commission too at times fell into selecting on the dependent variable. For instance, in a chapter recommending increased focus on character education, the commission notes that many school shooters experienced social isolation, without comparing this to levels of social isolation among those who do not engage in violence:

> In the aftermath of the Parkland shooting, multiple reports indicated the alleged shooter experienced feelings of isolation and depression in the years leading up to the shooting. . . . Perpetrators of previous school shootings shared that sense of detachment. For example, one Columbine shooter was characterized as depressed and reclusive. . . . Family members and acquaintances of the Virginia Tech shooter said that, as his isolation grew during his senior year, his "attention to schoolwork and class time dropped." . . . The same was true at Sandy Hook.

At times the commission does avoid selecting on the dependent variable. In a chapter on mental health, they write,

> Individuals who commit mass shootings may or may not have a serious mental illness (SMI). There is little population-level evidence to support the notion that those diagnosed with mental illness are more likely than anyone else to commit gun crimes.

But not long after, they return to arguments that suggest they are looking for correlation without variation:

> A U.S. Department of Education and U.S. Secret Service analysis found that as many as a quarter of individuals who committed mass shootings had been in treatment for mental illnesses. . . Such individuals often feel aggrieved and extremely angry, and nurture fantasies of violent revenge.

These are not the only such government reports; such analyses are seemingly inevitable after acts of youth violence. But, for reasons we've already seen, these findings, like the Senate testimonies above, are misleading. Even if it were true that virtually every young person who behaves in a troubling manner also listens to rock, reads comic books, or plays video games, this would not establish a correlation between such behavior and these supposed corrupters of the youth. Correlation requires variation. Evidence for the proposition that kids who engage in those activities are *more* likely to be violent than kids who do not engage in those activities must involve a comparison of these two types of kids.

If we want to know if there is a relationship between some putative scourge of the youth and violence, we must not select on the dependent variable—that is, we must compare violent kids to non-violent kids and see whether violent kids are more likely to engage in that scourge than non-violent kids. (Again, even then, we can't say the relationship is causal.) The fact that even experts can fail to think clearly about this means that, for all the expert opinion offered on the topic, we know far less than we could about the correlates of youth violence.

High School Dropouts

Let's stick, for the moment, with troubled youth. Early twenty-first-century America has a high school graduation problem. At a time when the economic returns to education are at an all-time high, almost a third of students in the public schools fail to complete high school on time. Over 10 percent never graduate.

In 2006, the Bill and Melinda Gates Foundation decided to put some resources into addressing this issue. As one step in trying to find a solution, they commissioned a study on the correlates of dropping out of high school. The report's main thrust is that high school dropout is not primarily associated with the things you might have guessed—problems at home, lack of academic preparation, or listening to rock music. Rather, the big problem seems to be that kids aren't engaged by the educational environment and find school boring.

As the report states, "nearly half (47 percent) [of dropouts] said a major reason for dropping out was that classes were not interesting." And "nearly 7 in 10 respondents (69 percent) said they were not motivated or inspired to work hard."

Unfortunately, because correlation requires variation, the evidence in this Gates Foundation study, just like the evidence presented by the PMRC and the anti–comic book lobby before it, is pretty uninformative.

The fact that half of high school dropouts report finding school uninteresting does not mean that finding school uninteresting correlates with dropout. Because correlation requires variation, measuring the correlation has to involve comparing dropouts to non-dropouts to see whether dropouts are more likely to find school uninteresting. The Gates Foundation study, because it looks only at high school dropouts, can't make this comparison.

This point isn't just pedantic. Think about it for a second. Both authors of this book went to high school. Neither dropped out. However, both authors recall finding some classes uninteresting. Didn't you?

Now, our personal experiences also don't constitute compelling evidence. So let's see if we can do a little better in figuring out whether finding classes boring is really a key predictor of dropout. Researchers at Indiana University did a nationally representative survey of high school students in 2009. Most of these students are not going to drop out, yet the researchers report that "two out of three respondents (66%) in 2009 are bored

at least every day in class." That's even more than the 50 percent of dropouts who find school boring in the Gates Foundation study.

But let's be careful. There are many reasons the Gates Foundation survey and the Indiana University survey can't be compared. They sample different groups of students, ask different questions, and are from different years. So we don't want to leap to conclusions. But at the very least, the Indiana University survey should make you worry that finding school boring is in fact a very common experience for high school students, not just those who drop out.

The future of American education is serious stuff. It is admirable that the Gates Foundation is trying to improve education. But their research ignores a key principle of thinking clearly with data; they are trying to learn about the correlates of educational failure without any variation in failure versus success. This approach cannot work.

Suicide Attacks

In 2009, University of Chicago professor and noted terrorism expert Robert Pape testified to the House of Representatives Armed Services Subcommittee on Terrorism. The topic was General Stanley McChrystal's proposal for a forty-thousand-troop surge to fight the Taliban insurgency in Afghanistan. Here is what Pape had to say:

> The picture is clear, the more Western troops have gone to Afghanistan, the more local residents have viewed themselves as under foreign occupation—and are using suicide and other terrorism to resist it. . .As my study of suicide terrorism around the world since 1980 shows, what motivates suicide terrorists is not the existence of a terrorist sanctuary, but the presence of foreign forces on land they prize. So, it is little surprise that US troops are producing anti-American suicide attackers.

Pape goes on to recommend a major rethinking of American military strategy in Afghanistan. His argument is based on the claim that suicide attacks are primarily motivated by foreign occupation. His evidence is the data he collected and analyzed in articles and two books on every suicide terrorist campaign in the world since 1980.

The argument sounds plausible. In Afghanistan, U.S. forces were being attacked by suicide bombers who wanted the United States to leave the country. Tamil Tiger suicide bombers attacked a government in Sri Lanka they believed was occupying their homeland. Palestinian suicide bombers attack Israelis, arguing that they are foreign occupiers. It sure seems like occupation is a major correlate of suicide attacks.

Now, the claim that virtually every suicide attack is targeted against a foreign occupier is, we think, debatable. (For instance, while Osama bin Laden claimed the American troops stationed in Saudi Arabia at the invitation of the Saudi government were an occupying force, are we sure we agree with him?) But, for the sake of argument, let's assume that the basic factual claim is correct. Does this mean that there is a correlation between foreign occupation and suicide attacks?

The answer is, of course, no. Correlation requires variation. To understand the correlates of suicide attacks, you can't just study every single instance of a suicide attack and look for commonalities. That is selecting on the dependent variable. You must compare conflicts with suicide attacks to those without.

An easy thing to do in this case is to simply look at every single country and ask: Are foreign-occupied countries more likely to experience suicide attacks than countries

that are not foreign occupied? It turns out that a recent study did precisely that comparison and found that the answer was no. In particular, if we compare occupied to non-occupied countries, the difference in likelihood of experiencing suicide violence is less than 1 percentage point!

What is going on? All those examples of suicide bombers that we listed involved attacking foreign occupiers. How could it be that there is almost no correlation between foreign occupation and suicide attacks?

The way to get some intuition is to think about how many foreign occupations there have been that didn't lead to suicide terrorism. The British occupation of Ireland, despite sparking a decades-long campaign of violent resistance, never gave rise to suicide terrorism. Basque separatists in Spain fought a decades-long campaign and never resorted to suicide attacks. At various points during the Cold War (and beyond), the United States stationed troops in Germany, Japan, South Korea, Grenada, Panama, and Haiti (arguably, all as much occupations as the putative occupation of Saudi Arabia) but suffered not even one suicide attack in any of these locations. If occupation predicts suicide violence, what was going on in all these places?

This example has another nice feature. It not only illustrates the mistake of looking for correlation without variation. It shows you how misled you can be by trying to reach conclusions by only looking at cases where the phenomenon of interest (here, suicide attacks) occurs—that is, by selecting on the dependent variable. To see this, it helps to go back in history a little.

Suppose you'd started collecting data on suicide violence in the early 1980s. By 1986 you'd have recorded thirty-three attacks and over one thousand deaths. Essentially every single one of those attacks was carried out by the armed Shi'a militia Hezbollah against American, Israeli, and French targets in Lebanon, including the attack on the U.S. Marines Barracks in Beirut, which killed 320 people.

If you'd looked for commonalities amongst every suicide attack ever committed in 1986, you might have noticed that they were all carried out by Muslims in the Middle East. Using the same logic that led to the conclusion that occupation is a major predictor of suicide attacks, you might have concluded that Islam was the key correlate.

Of course, if you had done a proper comparison, you wouldn't have reached this conclusion. There are a whole lot of Muslim-majority countries in the world. In 1986, almost none of them had experienced suicide violence.

Moreover, if you were trying to forecast where the next suicide attack might occur, this conclusion in 1986 would have led you terribly astray. In 1987, the world saw the first suicide attack by the Liberation Tigers of Tamil Eelam (Tamil Tigers), a group of secular separatists in Sri Lanka with no ties to Islam. The attack marked the beginning of what would become the largest campaign of suicide violence the world had ever seen. When you try to establish correlation without variation, you can get things colossally wrong.

The World Is Organized to Make Us Select on the Dependent Variable

As we've seen, it is incredibly easy to fall into the trap of selecting on the dependent variable simply by failing to think clearly. But matters are even worse than that. The world sometimes seems to be organized in a way that almost forces us to look for correlation without variation. In this section we look at three ways in which that is true:

the organization of certain professions, the practice of post-mortem analyses following disasters, and the way we seek life advice.

Doctors Mostly See Sick People

Anyone who has suffered from significant back pain knows that it is rough. When, inevitably, many of you develop back pain, you will likely go to a doctor, who will send you to get an MRI. Usually, the MRI shows some bulging or herniated discs in the afflicted back. These bulging discs are taken to be the cause, in some not fully understood way, of the back pain (maybe by impinging a nerve).

The recommendations following this diagnosis can vary greatly. Some doctors want to operate. Others will refer you to a pain clinic where yet other doctors might stick you with giant needles with medication that dulls pain and reduces inflammation. Still others will suggest you try physical therapy and take lots of painkillers.

Here's the kicker. As best we can tell, there is precious little evidence that having a bulging disc is correlated with back pain. Here are the facts. People with back pain are quite likely to exhibit disc herniation. Indeed, in a 2011 British study published in the journal *Pain*, about two-thirds of back pain sufferers who were referred for an MRI had nerve compression as a result of a disc bulge or herniation. This seems like evidence that those bulging discs really are a problem.

But remember, correlation requires variation. You should be asking yourself: What about people without back pain? How do their discs look? Good question. The answer is, they look exactly the same as the people's discs who do have back pain! A 1994 study published in the *New England Journal of Medicine* found that about two-thirds of people who do not suffer from back pain also have a disc bulge or herniation. Once you compare both variables of interest, the apparent association between bulging discs and back pain disappears.

It is easy to see how doctors could end up associating bulging discs with back pain. Even if they are thinking clearly, by dint of profession, a doctor is almost doomed not to look at variation. Sick people go to the doctor. Healthy people tend not to. Your typical back doctor just doesn't get much of an opportunity to look at the MRIs of people with well-functioning backs.

Post-Mortems

Another way the world is organized to make us look for correlation without variation is through institutional rules or procedures. A particularly common example is the way organizations respond to both great failures and great successes.

Following a crisis or disaster, organizations want to know what went wrong so they can avoid making similar mistakes in the future. Likewise, following great successes they want to know what went right to establish best practices. Achieving these goals is the role of a post-mortem analysis. Looking closely at an instance of great failure or great success is not, in and of itself, a mistake. Indeed, it is a very sensible starting point. But, if you think clearly, you should already be able to see that, on their own, such post-mortem procedures are not sufficient to establish correlations between what went wrong (or right) and existing practices.

The question you should be trying to answer when assessing lessons learned from a crisis is, Which decisions should have been made differently to avoid the crisis, given what we knew at the time? However, when assessing lessons learned, we often slip

Table 4.5. Rehearsal strategies in the week before competitions where your band performed poorly (made-up data).

	Do Well	Do Poorly	Total
Extra Rehearsals	?	80	?
Take It Easy	?	8	?
Total	?	88	?

into answering a slightly different question: Which decisions should have been made differently to avoid the crisis, given what we know now?

The latter isn't a terribly useful question to answer, for the reasons we've already talked about in this chapter. Suppose you find some decision that, it turns out, seems to have led directly to the disaster. After the fact, it is easy to say, "Had we not taken that action, the disaster wouldn't have happened." But does that mean that you shouldn't take similar such actions in the future? To know the answer to that, you'd want to know whether disasters are more likely to occur in the presence of such actions than in their absence. That is, you want to know whether there is a correlation between taking such actions and disasters occurring. To establish a correlation, you need variation. But a post-mortem, almost by definition, has no variation. You are only looking at an instance of the disaster occurring.

To see what we mean a little more intuitively, let's start with a fictional example. Then we'll turn to some real cases.

Imagine you are a high school band director preparing for a regional competition in a week. You have to decide whether to push the kids hard with a grueling schedule of rehearsals or give them time off so they go into the competition relaxed. You weigh the pros and cons, deciding preparation is more important than mental state. So you schedule a week of extra rehearsals. Unfortunately, the band doesn't play terribly well on the day of the competition, and you are eliminated in the first round.

In your post-mortem analysis you ask the question, What should I have done to avoid the loss? It occurs to you that you've seen a lot of bands lose competitions in this same way (i.e., having rehearsed themselves to death the week before), so you decide to collect some data. You look at the history of all the competitions in which your band was eliminated in the early rounds. Just like in this year's competition, you find that in almost every one of these competitions, you scheduled a heavy rehearsal schedule in the week leading up to the competition.

Let's say you did a week of intensive rehearsing prior to 80 out of 88 losses. The post-mortem conclusion seems clear. In over 90 percent of the cases where your band was eliminated early, it was after a week of exhausting rehearsal. Now you feel even more sure: intensive rehearsal is the wrong strategy. Table 4.5 summarizes what you know so far from your post-mortem analysis.

But this conclusion doesn't necessarily follow from the data you've collected. In fact, from this data alone, there's no way to know whether those rehearsals are associated with performing well or poorly, because you have answered the wrong question.

You don't want to know if bands did extra rehearsals prior to most of the competitions where they performed poorly. You want to know if extra rehearsals are positively or negatively correlated with performing well. The answer to that question will help you know whether those extra rehearsals are a good idea for the next competition.

Table 4.6. Rehearsal strategy in the week before competitions where your band performed well or poorly (made-up data).

	Do Well	Do Poorly	Total
Extra Rehearsals	300	80	380
Take It Easy	12	8	20
Total	312	88	400

To answer that question, you have to look at the correlation between extra rehearsals and performing well in competition. But you can't know the correlation from your post-mortem analysis. Correlation requires variation. Your post-mortem, by focusing only on poor performances, guarantees that you lack the variation needed to establish a correlation.

To do a better job, you could look at the history of all the band competitions you've participated in to see whether you performed well or poorly. Now you have variation in both variables and can fill in all the data, as shown in table 4.6.

From this table it is clear that there is in fact a strong positive correlation between scheduling extra rehearsals and performing well. The probability of your band performing well when you rehearsed hard is about 79 percent ($\frac{300}{380} \approx .79$). By contrast, the probability of your band performing well when you took it easy the week prior to a competition is only 60 percent ($\frac{12}{20} = .60$). The only reason that the post-mortem turned up the finding that almost every poor performance involved intensive rehearsals is that those extra rehearsals are so effective that sensible band directors almost always schedule them.

By finding the variation needed to establish the correlation that is actually relevant to the question at hand, you reach a very different conclusion than you did in your original post-mortem. Following the loss, it seemed like intensive rehearsals were a bad idea. But before the fact, given the information available, rehearsing hard was exactly the right call. Faced with the same situation again, you should probably make the same decision.

This problem is endemic to the process of post-mortems following disasters. We tend to look at the factors that seem like they contributed to the disaster, ask if they were also present in past disasters, and, when they were, conclude that we should eliminate those factors in the future. But, in so doing, we are making the same mistake as the band director. Without variation in whether or not a disaster occurred, we can't actually learn whether the presence of those factors is correlated with the occurrence of a disaster. So we don't know if there are lessons to be learned.

We are going to show you what we mean with two examples of post-mortems that followed major disasters—the *Challenger* space shuttle explosion in 1986 and the financial crisis of 2008. In each case, we will see that, while after the fact it sure looks like some serious and obvious mistakes were made, it is less clear that the decision makers could have known that they were making mistakes before the fact. Moreover, once we've grasped this, we will be able to think more clearly about how to design post-mortems that might be more informative about lessons learned.

The Challenger *disaster*

On January 28, 1986, the space shuttle *Challenger* disintegrated off the coast of Cape Canaveral less than two minutes after launch. Seven crew members were killed. The

night before the *Challenger* exploded, a small group of engineers from the NASA contractor responsible for the shuttle's solid rocket boosters predicted that the cold weather would lead to a catastrophic failure that might well compromise the shuttle. The concern was that the critical O-ring seals responsible for containing gases produced by burning rocket fuel were not certified to operate at the low temperatures that preceded this particular launch. If the O-ring seals failed, the engineers argued, hot pressurized gas could burn through the rocket's casing, causing disaster.

These predictions, shunted aside by managers at NASA and the engineers' own firm, proved tragically correct. Many post-mortem analyses focused on NASA's failure to take these concerns seriously. The conclusion most observers reached was that the disaster was caused by organizational and cultural failures at NASA that facilitated groupthink and led managers to systematically ignore important objections from experts. For instance, the *Report of the Presidential Commission on the Space Shuttle Challenger Accident* (the Rogers Commission) concluded, "Failures in communication . . . resulted in a decision to launch 51-L based on incomplete and sometimes misleading information, a conflict between engineering data and management judgements, and a NASA management structure that permitted internal flight safety problems to bypass key Shuttle managers."

The *Challenger* case is interesting. No one questions the physics behind the conclusion that the O-rings failed because of cold temperatures. Indeed, the Rogers Commission included the Nobel Prize–winning physicist Richard Feynman precisely so they could say with authority whether the engineers were right on the science. They were. And so, in this sense, launching the shuttle was clearly a mistake.

Because the science is so clear, it seems natural for a post-mortem to ask what it was about the process that led decision makers to ignore engineers making good scientific arguments. Here is where our knowledge of the pitfalls of post-mortems should make us stop and think. We know that, after the fact, the decision to launch was tragically flawed. But we want to evaluate whether it was a bad decision at the time it was made. To do so, we need to know about the correlation between the presence of scientifically valid engineering concerns and the success of shuttle launches. And to know about that correlation, we need variation; we must compare disastrous launches to successful launches.

We aren't engineers, so we aren't going to try to weigh in on whether or not the decision to launch *Challenger* was reasonable at the time it was made. But we can see how, to analyze this, a post-mortem commission would need to ask questions they aren't accustomed to asking. Post-mortem commissions ask what led to the disaster, whether people had raised the relevant objections, and, if so, why those objections weren't listened to. In addition, such commissions need to ask whether engineers also raised scientifically valid concerns prior to lots of successful launches. This doesn't seem implausible. Space shuttle launches are incredibly complex and dangerous undertakings. Perhaps there is almost always a scientifically valid reason for serious concern. If so, then there actually wouldn't be much (if any) correlation between the presence of such concerns and launch success. If this is the case, unless you are prepared to simply shut down the space program, it isn't fair to say that launching following a scientifically plausible objection by an engineer is always a mistake. This is the sort of thing one would want to know from a post-mortem commission before reaching conclusions about changing NASA's organizational culture or management practices.

The financial crisis of 2008

The financial crisis that shook the world economy in 2007 and 2008 began with a crash in the U.S. subprime housing market. This crash had ripple effects across the banking sector that eventually spread throughout the world. Understandably, in the wake of this crisis—at the time, the worst since the Great Depression—policy makers and the public alike were interested in identifying early warning indicators that might help them forecast and forestall future crises.

Perhaps the most important post-mortem analysis attempting to provide such early warning indicators was the book *This Time Is Different* by the economists Carmen M. Reinhart and Kenneth S. Rogoff. Reinhart and Rogoff collected and analyzed data on every major financial crisis of the last eight hundred years. By doing so, they argued, they could identify a few key indicators that almost always precede such a crisis. These include uncommonly large current account deficits (that is, goods and services exported minus imported net of income from abroad), asset price bubbles, and excessive borrowing. For instance, in 2006 the United States had a current account deficit close to 7 percent of GDP, a bubble in the housing market, and ballooning federal debt. Thus, Reinhart and Rogoff conclude, "we've been here before." The implication is that the 2008 U.S. financial crisis could have been predicted by the presence of those same factors that seem to characterize financial crises across time and around the globe. Similar patterns were true before the financial crises in Latin America in the early 2000s, East Asia in the 1990s, Nordic countries in the 1980s, and so on into history.

The problem with this argument is the same as in our earlier examples. Early warning indicators should be correlates of financial crises. Because correlation requires variation, to know if current account deficits, soaring asset prices, and heavy borrowing correlate with financial crises, we need variation in crises. That is, we need to know not only that these factors tend to be present when crises occur but also how frequently they are present when crises do not occur. Without such variation, we cannot establish a correlation.

Reinhart and Rogoff's plan of studying every major financial crisis for eight hundred years cannot answer the question. And there are reasons to be worried about their conclusions. As the MIT political scientist David Andrew Singer points out, one need only look at recent history to cast some doubt on the story. For instance, in the late 1990s the United States had all the early warning signs for a financial crisis. There were large current account deficits as a result of massive foreign investment in dot-coms. Moreover, when the dot-com bubble burst, "it wiped out approximately $5 trillion in market capitalization." Yet no financial crisis occurred. This, of course, is just one anecdote. But it should make you wonder whether the factors Reinhart and Rogoff point to are really good predictors of financial crises or just common features of the world that happen to exist both when financial crises occur and when they don't occur.

Life Advice

We've been arguing that our world is organized in ways that lead us to try to figure out the correlates of success or failure without looking at variation, even though it won't work. It is important to see that this problem isn't confined to big institutional settings. We are all victims of it every day in many small ways.

One simple example is the ways in which we seek life advice, which almost always involves asking successful people how it is that they succeeded. In our business, for instance, graduate students are encouraged to ask senior professors what they did to succeed on the job market. We imagine something similar is true in other professions. There is certainly no shortage of self-help books describing the habits of successful people.

But such wisdom suffers from exactly the problems we've been pointing to. Successful people, reflecting on their lives, are inclined to identify a few decisions they made or a few personal characteristics that seem important and offer them as advice to the next generation. But those successful people typically have no idea whether many other, less successful people made similar decisions or had similar characteristics. That is, their introspection about the correlates of success lacks variation. As such, successful people don't really know whether the lessons they point to in telling their personal stories are correlates of success or not. And so, we leave you with this happy bit of wisdom of our own: Beware life advice. Most of it is probably nonsense.

Wrapping Up

Correlation requires variation. But unclear thinking and organizational mandates often lead us to select on the dependent variable—trying to establish the correlates of some phenomenon by only looking at instances when it occurred. It requires careful attention to make sure you aren't falling into this trap, whether you are doing quantitative analysis or just trying to think informally about evidence. Even just forcing yourself to think about whether you could fill in all four cells of one of our two-by-two tables is a good starting point for avoiding looking for correlation without variation.

You can be even more rigorous by using quantitative techniques to measure correlations. The most important such technique is called regression, the topic of chapter 5.

Key Term

- **Selecting on the dependent variable:** Examining only instances when the phenomenon of interest occurred, rather than comparing cases where it occurred to cases where it did not occur.

Exercises

4.1 In chapter 2 we discussed the differences between statements about correlations and other factual statements that do not convey information about a correlation. Now that you have a deeper understanding that correlation requires variation, consider the following statements. Which ones describe a correlation, and which ones do not?

(a) Most top-performing schools have small student bodies.
(b) Married people are typically happier than unmarried people.
(c) Among professionals, taller basketball players tend to have lower free-throw percentages than shorter players.
(d) The locations in the United States with the highest cancer rates are typically small towns.
(e) Older houses are more likely to have lead paint than newer ones.

(f) Most colds caught in Cook County are caught on cold days. (This one also doubles as a tongue twister.)

4.2 At least twenty billionaires dropped out of college before earning their fortunes, including Bill Gates and Mark Zuckerberg.

(a) Does this mean that dropping out of college is correlated with becoming a billionaire? Why or why not?
(b) Draw the two-by-two table that would allow you to assess whether dropping out of college is correlated with becoming a billionaire. Let's assume that exactly twenty people have dropped out of college and become billionaires, so you know what to put in one of the four cells. Make your best guess for the other cells. At the time of this writing, there are about 7.8 billion people in the world, and about two thousand billionaires. Do you think there is a positive or negative correlation between dropping out of college and becoming a billionaire?
(c) Given your guesses from part (b), what proportion of the non-billionaires would need to be college dropouts in order for the correlation to be negative? What proportion of the non-billionaires would need to be college dropouts in order for the correlation to be positive?
(d) If you're currently a college student deciding whether you want to drop out in the hopes of becoming a billionaire, you may want to restrict attention to people who actually started college. Do you think the correlation between dropping out of college and becoming a billionaire is more or less likely to be positive if we restrict attention to just people who start college?
(e) About 7 percent of the world's population has a college degree. And about a third of people who start college complete it. If we assume that everyone who becomes a billionaire started college, you should now have all the information you need to assess the correlation between becoming a billionaire and dropping out of college among those who start college. Is it positive, negative, or zero?

4.3 Identify one recent case where an analyst made the mistake discussed in this chapter. That is, find a case where someone (at least implicitly) makes a claim about a correlation but they don't have variation in one of their variables. Your example might come from a newspaper article, an academic study, a policy memo, or a statement from a politician or business leader.

(a) Summarize the claim being made (perhaps implicitly) and explain why the evidence does not necessarily support the claim.
(b) Explain what additional data collection and analysis *would* allow the analyst to assess the correlation of interest.
(c) Draw a two-by-two table that illustrates your argument, and discuss what the unknown numbers in the table would have to be in order for the correlation of interest to be positive, negative, or zero.