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—$Y_{\text{college}}$ 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/.