9

Causality

Our starting point is the difference between an observation and an action. What we see in passive observation is how individuals follow their routine behavior, habits, and natural inclination. Passive observation reflects the state of the world projected to a set of features we chose to highlight. Data that we collect from passive observation show a snapshot of our world as it is.

There are many questions we can answer from passive observation alone: Do 16 year-old drivers have a higher incidence rate of traffic accidents than 18 year-old drivers? Formally, the answer corresponds to a difference of conditional probabilities. We can calculate the conditional probability of a traffic accident given that the driver's age is 16 years and subtract from it the conditional probability of a traffic accident given the age is 18 years. Both conditional probabilities can be estimated from a large enough sample drawn from the distribution, assuming that there are both 16 year old and 18 year old drivers. The answer to the question we asked is solidly in the realm of observational statistics.

But important questions often are not observational in nature. Would traffic fatalities decrease if we raised the legal driving age by two years? Although the question seems similar on the surface, we quickly realize that it asks for a fundamentally different insight. Rather than asking for the frequency of an event in our manifested world, this question asks for the effect of a hypothetical action.

As a result, the answer is not so simple. Even if older drivers have a lower incidence rate of traffic accidents, this might simply be a consequence of additional driving experience. There is no obvious reason why an 18 year old with two months on the road would be any less likely to be involved in an accident than, say, a 16 year-old with the same experience. We can try to address this problem by holding the number of months of driving experience fixed, while comparing individuals of different ages. But we quickly run into subtleties. What if 18 year-olds with two months of driving experience correspond to individuals who are exceptionally cautious and hence—by their natural inclination—not only drive less, but also more cautiously? What if such individuals predominantly live in regions where traffic conditions differ significantly from those in areas where people feel a

greater need to drive at a younger age?

We can think of numerous other strategies to answer the original question of whether raising the legal driving age reduces traffic accidents. We could compare countries with different legal driving ages, say, the United States and Germany. But again, these countries differ in many other possibly relevant ways, such as, the legal drinking age.

At the outset, causal reasoning is a conceptual and technical framework for addressing questions about the effect of hypothetical actions or *interventions*. Once we understand what the effect of an action is, we can turn the question around and ask what action plausibly *caused* an event. This gives us a formal language to talk about cause and effect.

The limitations of observation

Before we develop any new formalism, it is important to understand why we need it in the first place.

To see why we turn to the venerable example of graduate admissions at the University of California, Berkeley in 1973.[1] Historical data show that 12763 applicants were considered for admission to one of 101 departments and inter-departmental majors. Of the 4321 women who applied roughly 35 percent were admitted, while 44 percent of the 8442 men who applied were admitted. Standard statistical significance tests suggest that the observed difference would be highly unlikely to be the outcome of sample fluctuation if there were no difference in underlying acceptance rates.

A similar pattern exists if we look at the aggregate admission decisions of the six largest departments. The acceptance rate across all six departments for men is about 44%, while it is only roughly 30% for women, again, a significant difference. Recognizing that departments have autonomy over who to admit, we can look at the gender bias of each department.

Table 1: UC Berkeley admissions data from 1973.

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

What we can see from the table is that four of the six largest departments show a higher acceptance ratio among women, while two show a higher acceptance rate for men. However, these two departments cannot account for the large difference in acceptance rates that we observed in aggregate. So, it appears that the higher acceptance rate for men that we observed in aggregate seems to have reversed at the department level.

Such reversals are sometimes called *Simpson's paradox*, even though mathematically they are no surprise. It's a fact of conditional probability that there can be events Y (here, acceptance), A (here, female gender taken to be a binary variable) and a random variable Z (here, department choice) such that:

1. $\mathbb{P}[Y \mid A] < \mathbb{P}[Y \mid \neg A]$
2. $\mathbb{P}[Y \mid A, Z = z] > \mathbb{P}[Y \mid \neg A, Z = z]$ for all values z that the random variable Z assumes.

Simpson's paradox nonetheless causes discomfort to some, because intuition suggests that a trend which holds for all subpopulations should also hold at the population level.

The reason why Simpson's paradox is relevant to our discussion is that it's a consequence of how we tend to misinterpret what information conditional probabilities encode. Recall that a statement of conditional probability corresponds to passive observation. What we see here is a snapshot of the normal behavior of women and men applying to graduate school at UC Berkeley in 1973.

What is evident from the data is that gender influences department choice. Women and men appear to have different preferences for different fields of study. Moreover, different departments have different admission criteria. Some have lower acceptance rates, some higher. Therefore, one explanation for the data we see is that women *chose* to apply to more competitive departments, hence getting rejected at a higher rate than men.

Indeed, this is the conclusion the original study drew:

> *The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.*[1]

In other words, the article concluded that the source of gender bias in admissions was a *pipeline problem*: Without any wrongdoing by the

departments, women were "shunted by their socialization" that happened at an earlier stage in their lives.

It is difficult to debate this conclusion on the basis of the available data alone. The question of discrimination, however, is far from resolved. We can ask why women applied to more competitive departments in the first place. There are several possible reasons. Perhaps less competitive departments, such as engineering schools, were unwelcoming of women at the time. This may have been a general pattern at the time or specific to the university. Perhaps some departments had a track record of poor treatment of women that was known to the applicants. Perhaps the department advertised the program in a manner that discouraged women from applying.

The data we have also shows no measurement of *qualification* of an applicant. It's possible that due to self-selection women applying to engineering schools in 1973 were over-qualified relative to their peers. In this case, an equal acceptance rate between men and women might actually be a sign of discrimination.

There is no way of knowing what was the case from the data we have. We see that at best the original analysis leads to a number of follow-up questions.

At this point, we have two choices. One is to design a new study and collect more data in a manner that might lead to a more conclusive outcome. The other is to argue over which scenario is more likely based on our beliefs and plausible assumptions about the world.

Causal inference is helpful in either case. On the one hand, it can be used as a guide in the design of new studies. It can help us choose which variables to include, which to exclude, and which to hold constant. On the other hand, causal models can serve as a mechanism to incorporate scientific domain knowledge and exchange plausible assumptions for plausible conclusions.

Causal models

We choose *structural causal models* as the basis of our formal discussion as they have the advantage of giving a sound foundation for various causal notions we will encounter. The easiest way to conceptualize a structural causal model is as a program for generating a distribution from independent noise variables through a sequence of formal instructions. Imagine instead of samples from a distribution, somebody gave you a step-by-step computer program to generate samples on your own starting from a random seed. The process is not unlike how you would write code. You start from a simple random seed and build up increasingly more complex constructs. That is basically what a structural causal model is, except that each assignment

uses the language of mathematics rather than any concrete programming syntax.

A first example

Let's start with a toy example not intended to capture the real world. Imagine a hypothetical population in which an individual exercises regularly with probability $1/2$. With probability $1/3$, the individual has a latent disposition to develop overweight that manifests in the absence of regular exercise. Similarly, in the absence of exercise, heart disease occurs with probability $1/3$. Denote by X the indicator variable of regular exercise, by W that of excessive weight, and by H the indicator of heart disease. Below is a structural causal model to generate samples from this hypothetical population. Recall a Bernoulli random variable $\mathrm{B}(p)$ with bias p is a biased coin toss that assumes value 1 with probability p and value 0 with probability $1 - p$.

1. Sample independent Bernoulli random variables $U_1 \sim \mathrm{B}(1/2)$, $U_2 \sim \mathrm{B}(1/3)$, $U_3 \sim \mathrm{B}(1/3)$.
2. $X := U_1$
3. $W :=$ if $X = 1$ then 0 else U_2
4. $H :=$ if $X = 1$ then 0 else U_3

Contrast this generative description of the population with a usual random sample drawn from the population that might look like this:

X	W	H
0	1	1
1	0	0
1	1	1
1	1	0
0	1	0
...	...	...

From the program description, we can immediately see that in our hypothetical population *exercise* averts both *overweight* and *heart disease*, but in the absence of exercise the two are independent. At the outset, our program generates a joint distribution over the random variables (X, W, H). We can calculate probabilities under this distribution. For example, the probability of heart disease under the distribution specified by our model is $1/2 \cdot 1/3 = 1/6$. We can also calculate the conditional probability of heart diseases given overweight. From the event $W = 1$ we can infer that the

individual does not exercise so that the probability of heart disease given overweight increases to $1/3$ compared with the baseline of $1/6$.

Does this mean that overweight causes heart disease in our model? The answer is *no* as is intuitive given the program to generate the distribution. But let's see how we would go about arguing this point formally. Having a program to generate a distribution is substantially more powerful than just having sampling access. One reason is that we can manipulate the program in whichever way we want, assuming we still end up with a valid program. We could, for example, set $W := 1$, resulting in a new distribution. The resulting program looks like this:

2. $X := U_1$
3. $W := 1$
4. $H :=$ if $X = 1$ then 0 else U_3

This new program specifies a new distribution. We can again calculate the probability of heart disease under this new distribution. We still get $1/6$. This simple calculation reveals a significant insight. The substitution $W := 1$ does not correspond to a conditioning on $W = 1$. One is an action, albeit inconsequential in this case. The other is an observation from which we can draw inferences. If we observe that an individual is overweight, we can infer that they have a higher risk of heart disease (in our toy example). However, this does not mean that lowering body weight would avoid heart disease. It wouldn't in our example. The active substitution $W := 1$ in contrast creates a new hypothetical population in which all individuals are overweight with all that it entails in our model.

Let us belabor this point a bit more by considering another hypothetical population, specified by the equations:

2. $W := U_2$
3. $X :=$ if $W = 0$ then 0 else U_1
4. $H :=$ if $X = 1$ then 0 else U_3

In this population exercise habits are driven by body weight. Overweight individuals choose to exercise with some probability, but that's the only reason anyone would exercise. Heart disease develops in the absence of exercise. The substitution $W := 1$ in this model leads to an increased probability of exercise, hence lowering the probability of heart disease. In this case, the conditioning on $W = 1$ has the same affect. Both lead to a probability of $1/6$.

What we see is that fixing a variable by substitution may or may not correspond to a conditional probability. This is a formal rendering of our earlier point that observation isn't action. A substitution corresponds to an

action we perform. By substituting a value we break the natural course of action our model captures. This is the reason why the substitution operation is sometimes called the *do-operator*, written as $\text{do}(W := 1)$.

Structural causal models give us a formal calculus to reason about the effect of hypothetical actions. We will see how this creates a formal basis for all the different causal notions that we will encounter in this chapter.

Structural causal models, more formally

Formally, a structural causal model is a sequence of assignments for generating a joint distribution starting from independent noise variables. By executing the sequence of assignments we incrementally build a set of jointly distributed random variables. A structural causal model therefore not only provides a joint distribution, but also a description of how the joint distribution can be generated from elementary noise variables. The formal definition is a bit cumbersome compared with the intuitive notion.

Definition 1. *A* structural causal model *M is given by a set of variables $X_1, ..., X_d$ and corresponding assignments of the form*

$$X_i := f_i(P_i, U_i), \qquad i = 1, ..., d\,.$$

Here, $P_i \subseteq \{X_1, ..., X_d\}$ is a subset of the variables that we call the parents *of X_i. The random variables $U_1, ..., U_d$ are called* noise variables*, which we require to be jointly independent.*

The directed graph corresponding to the model has one node for each variable X_i, which has incoming edges from all the parents P_i. We will call such a graph the causal graph *corresponding to the structural causal model.*

The noise variables that appear in the definition model *exogenous factors* that influence the system. Consider, for example, how the weather influences the delay on a traffic route you choose. Due to the difficulty of modeling the influence of weather more precisely, we could take the weather induced to delay to be an exogenous factor that enters the model as a noise variable. The choice of exogenous variables and their distribution can have important consequences for what conclusions we draw from a model.

The parent nodes P_i of node i in a structural causal model are often called the *direct causes* of X_i. Similarly, we call X_i the *direct effect* of its direct causes P_i. Recall our hypothetical population in which weight gain was determined by lack of exercise via the assignment $W := \min\{U_1, 1 - X\}$. Here we would say that exercise (or lack thereof) is a direct cause of weight gain.

Structural causal model are a collection of formal *assumptions* about how certain variables interact. Each assignment specifies a *response function*. We

can think of nodes as receiving messages from their parents and acting according to these messages as well as the influence of an exogenous noise variable.

To which extent a structural causal model conforms to reality is a separate and difficult question that we will return to in more detail later. For now, think of a structural causal model as formalizing and exposing a set of assumptions about a data generating process. As such different models can expose different hypothetical scenarios and serve as a basis for discussion. When we make statements about cause and effect in reference to a model, we don't mean to suggest that these relationship necessarily hold in the real world. Whether they do depends on the scope, purpose, and validity of our model, which may be difficult to substantiate.

It's not hard to show that a structural causal model defines a unique joint distribution over the variables $(X_1, ..., X_d)$ such that $X_i = f_i(P_i, U_i)$. It's convenient to introduce a notion for probabilities under this distribution. When M denotes a structural causal model, we will write the probability of an event E under the entailed joint distribution as $\mathbb{P}_M\{E\}$. To gain familiarity with the notation, let M denote the structural causal model for the hypothetical population in which both weight gain and heart disease are directly caused by an absence of exercise. We calculated earlier that the probability of heart disease in this model is $\mathbb{P}_M\{H\} = 1/6$.

In what follows we will derive from this single definition of a structural causal model all the different notions and terminology that we'll need in this chapter.

Throughout, we restrict our attention to acyclic assignments. Many real-world systems are naturally described as stateful dynamical system with feedback loops. For example, often cycles can be broken up by introducing time dependent variables, such as, investments at time 0 grow the economy at time 1 which in turn grows investments at time 2, continuing so forth until some chosen time horizon t. We will return to a deeper dive into dynamical systems and feedback in later chapters.

Causal graphs

We saw how structural causal models naturally give rise to *causal graphs* that represent the assignment structure of the model graphically. We can go the other way as well by simply looking at directed graphs as placeholders for an unspecified structural causal model which has the assignment structure given by the graph. Causal graphs are often called *causal diagrams*. We'll use these terms interchangeably.

Below we see causal graphs for the two hypothetical populations from

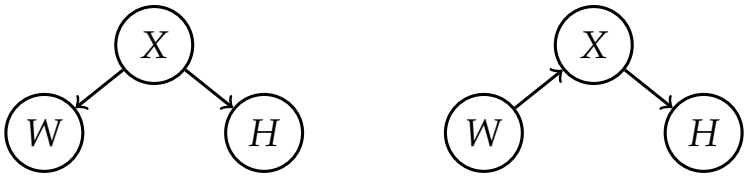

Figure 1: Causal diagrams for the heart disease examples.

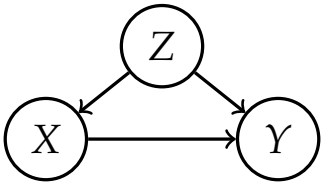

Figure 2: Example of a fork.

our heart disease example.

The scenarios differ in the direction of the link between exercise and weight gain.

Causal graphs are convenient when the exact assignments in a structural causal models are of secondary importance, but what matters are the paths present and absent in the graph. Graphs also let us import the established language of graph theory to discuss causal notions. We can say, for example, that an *indirect cause* of a node is any ancestor of the node in a given causal graph. In particular, causal graphs allow us to distinguish cause and effect based on whether a node is an ancestor or descendant of another node.

Let's take a first glimpse at a few important graph structures.

Forks

A *fork* is a node Z in a graph that has outgoing edges to two other variables X and Y. Put differently, the node Z is a common cause of X and Y.

We already saw an example of a fork in our weight and exercise example: $W \leftarrow X \rightarrow H$. Here, exercise X influences both weight and heart disease. We also learned from the example that Z has a *confounding* effect: Ignoring exercise X, we saw that W and H appear to be positively correlated. However, the correlation is a mere result of confounding. Once we hold exercise levels constant (via the do-operation), weight has no effect on heart disease in our example.

Confounding leads to a disagreement between the calculus of conditional probabilities (observation) and do-interventions (actions).

Real-world examples of confounding are a common threat to the validity of conclusions drawn from data. For example, in a well known medical

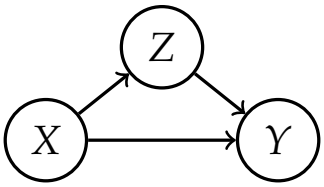

Figure 3: Example of a mediator.

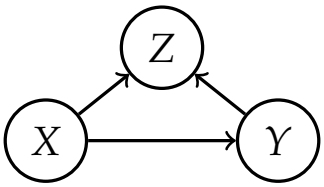

Figure 4: Example of a collider.

study a suspected beneficial effect of *hormone replacement therapy* in reducing cardiovascular disease disappeared after identifying *socioeconomic status* as a confounding variable.[2]

Mediators

The case of a fork is quite different from the situation where Z lies on a directed path from X to Y:

In this case, the path $X \to Z \to Y$ contributes to the total effect of X on Y. It's a causal path and thus one of the ways in which X causally influences Y. That's why Z is not a confounder. We call Z a *mediator* instead.

We saw a plausible example of a mediator in our UC Berkeley admissions example. In one plausible causal graph, department choice mediates the influences of gender on the admissions decision.

Colliders

Finally, let's consider another common situation: the case of a *collider*.

Colliders aren't confounders. In fact, in the above graph, X and Y are unconfounded, meaning that we can replace do-statements by conditional probabilities. However, something interesting happens when we condition on a collider. The conditioning step can create correlation between X and Y, a phenomenon called *explaining away*. A good example of the explaining away effect, or *collider bias*, is known as Berkson's paradox.[3] Two independent diseases can become negatively correlated when analyzing hospitalized patients. The reason is that when either disease (X or Y) is sufficient for admission to the hospital (indicated by variable Z), observing that a patient

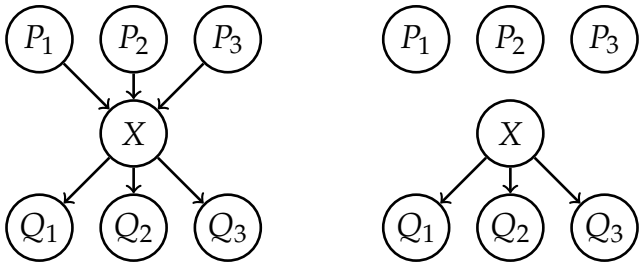

Figure 5: Graph before and after substitution.

has one disease makes the other statistically less likely. Berkson's paradox is a cautionary tale for statistical analysis when we're studying a cohort that has been subjected to a selection rule.

Interventions and causal effects

Structural causal models give us a way to formalize the effect of hypothetical actions or interventions on the population within the assumptions of our model. As we saw earlier all we needed was the ability to do substitutions.

Substitutions and the do-operator

Given a structural causal model M we can take any assignment of the form

$$X := f(P, U)$$

and replace it by another assignment. The most common substitution is to assign X a constant value x:

$$X := x$$

We will denote the resulting model by $M' = M[X := x]$ to indicate the surgery we performed on the original model M. Under this assignment we hold X constant by removing the influence of its parent nodes and thereby any other variables in the model.

Graphically, the operation corresponds to eliminating all incoming edges to the node X. The children of X in the graph now receive a fixed message x from X when they query the node's value.

The assignment operator is also called the *do-operator* to emphasize that it corresponds to performing an action or intervention. We already have notation to compute probabilities after applying the do-operator, namely, $\mathbb{P}_{M[X:=x]}(E)$.

Another notation is popular and common:

$$\mathbb{P}\{E \mid \mathrm{do}(X := x)\} = \mathbb{P}_{M[X:=x]}(E)$$

This notation analogizes the do-operation with the usual notation for conditional probabilities, and is often convenient when doing calculations involving the do-operator. Keep in mind, however, that the do-operator (action) is fundamentally different from the conditioning operator (observation).

Causal effects

The *causal effect* of an action $X := x$ on a variable Y refers to the distribution of the variable Y in the model $M[X := x]$. When we speak of the causal effect of a variable X on another variable Y we refer to all the ways in which setting X to any possible value x affects the distribution of Y.

Often times X denotes the presence or absence of an intervention or *treatment*. In such case, X is a binary variable and are interested in a quantity such as

$$\mathbb{E}_{M[X:=1]}[Y] - \mathbb{E}_{M[X:=0]}[Y]\,.$$

This quantity is called the *average treatment effect*. It tells us how much treatment (action $X := 1$) increases the expectation of Y relative to no treatment (action $X := 0$).

Causal effects are population quantities. They refer to effects averaged over the whole population. Often the effect of treatment varies greatly from one individual or group of individuals to another. Such treatment effects are called *heterogeneous*.

Confounding

Important questions in causality relate to when we can rewrite a do-operation in terms of conditional probabilities. When this is possible, we can estimate the effect of the do-operation from conventional conditional probabilities that we can estimate from data.

The simplest question of this kind asks when a causal effect $\mathbb{P}\{Y = y \mid \mathrm{do}(X := x)\}$ coincides with the condition probability $\mathbb{P}\{Y = y \mid X = x\}$. In general, this is not true. After all, the difference between observation (conditional probability) and action (interventional calculus) is what motivated the development of causality.

The disagreement between interventional statements and conditional statements is so important that it has a well-known name: *confounding*. We

say that X and Y are confounded when the causal effect of action $X := x$ on Y does not coincide with the corresponding conditional probability.

When X and Y are confounded, we can ask if there is some combination of conditional probability statements that give us the desired effect of a do-intervention. This is generally possible given a causal graph by conditioning on the parent nodes PA of the node X:

$$\mathbb{P}\{Y = y \mid \mathrm{do}(X := x)\} = \sum_z \mathbb{P}\{Y = y \mid X = x, PA = z\}\mathbb{P}\{PA = z\}$$

This formula is called the *adjustment formula*. It gives us one way of estimating the effect of a do-intervention in terms of conditional probabilities.

The adjustment formula is one example of what is often called *controlling for* a set of variables: We estimate the effect of X on Y separately in every slice of the population defined by a condition $Z = z$ for every possible value of z. We then average these estimated sub-population effects weighted by the probability of $Z = z$ in the population. To give an example, when we control for age, we mean that we estimate an effect separately in each possible age group and then average out the results so that each age group is weighted by the fraction of the population that falls into the age group.

Controlling for more variables in a study isn't always the right choice. It depends on the graph structure. Let's consider what happens when we control for the variable Z in the three causal graphs we discussed above.

- Controlling for a confounding variable Z in a fork $X \leftarrow Z \rightarrow Y$ will deconfound the effect of X on Y.
- Controlling for a mediator Z will eliminate some of the causal influence of X on Y.
- Controlling for a collider will create correlation between X and Y. That is the opposite of what controlling for Z accomplishes in the case of a fork. The same is true if we control for a descendant of a collider.

The backdoor criterion

At this point, we might worry that things get increasingly complicated. As we introduce more nodes in our graph, we might fear a combinatorial explosion of possible scenarios to discuss. Fortunately, there are simple sufficient criteria for choosing a set of deconfounding variables that is safe to control for.

A well known graph-theoretic notion is the *backdoor* criterion.[4] Two variables are confounded if there is a so-called *backdoor* path between them. A *backdoor path* from X to Y is any path starting at X with a backward edge "$\leftarrow$" into X such as:

$$X \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$$

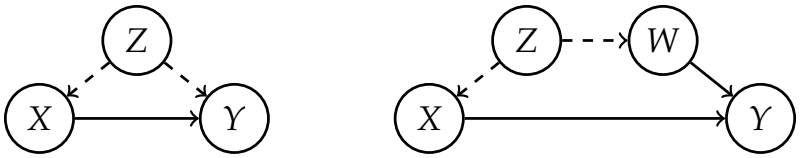

Figure 6: Two cases of unobserved confounding.

Intuitively, backdoor paths allow information flow from X to Y in a way that is not causal. To deconfound a pair of variables we need to select a *backdoor set* of variables that "blocks" all backdoor paths between the two nodes. A backdoor path involving a chain $A \to B \to C$ can be blocked by controlling for B. Information by default cannot flow through a collider $A \to B \leftarrow C$. So we only have to be careful not to open information flow through a collider by conditioning on the collider, or descendant of a collider.

Unobserved confounding

The adjustment formula might suggest that we can always eliminate confounding bias by conditioning on the parent nodes. However, this is only true in the absence of *unobserved confounding*. In practice often there are variables that are hard to measure, or were simply left unrecorded. We can still include such unobserved nodes in a graph, typically denoting their influence with dashed lines, instead of solid lines.

The above figure shows two cases of unobserved confounding. In the first example, the causal effect of X on Y is unidentifiable. In the second case, we can block the confounding backdoor path $X \leftarrow Z \to W \to Y$ by controlling for W even though Z is not observed. The backdoor criterion lets us work around unobserved confounders in some cases where the adjustment formula alone wouldn't suffice.

Unobserved confounding nonetheless remains a major obstacle in practice. The issue is not just lack of measurement, but often lack of anticipation or awareness of a confounding variable. We can try to combat unobserved confounding by increasing the number of variables under consideration. But as we introduce more variables into our study, we also increase the burden of coming up with a valid causal model for all variables under consideration. In practice, it is not uncommon to control for as many variables as possible in a hope to disable confounding bias. However, as we saw, controlling for mediators or colliders can be harmful.

Randomization and the backdoor criterion

The backdoor criterion gives a non-experimental way of eliminating confounding bias given a causal model and a sufficient amount of observational data from the joint distribution of the variables. An alternative experimental method of eliminating confounding bias randomization.

The idea is simple. If a treatment variable T is an unbiased coin toss, nothing but mere chance influenced it assignment. In particular, there cannot be a confounding variable exercising influence on both the treatment variable and a desired outcome variable.

A different way to think about is that randomization breaks natural inclination. Rather than letting treatment take on its natural value, we assign it randomly. Thinking in terms of causal models, what this means is that we eliminate all incoming edges into the treatment variable. In particular, this closes all backdoor paths and hence avoids confounding bias. Because randomization is such an important part of causal inference, we now turn to it in greater detail.

Experimentation, randomization, potential outcomes

Let's think about experimentation from first principles. Suppose we have a population of individuals and we have devised some treatment that can be applied to each individual. We would like to know the effect of this treatment on some measurable quantity.

As a simple example, and one which has had great utility, consider the development of a vaccine for a disease. How can we tell if a vaccine prevents disease? If we give everyone the vaccine, we'd never be able to disentangle whether the treatment caused the associated change in disease we observe or not. The common and widely accepted solution in medicine is to restrict our attention to a subset of the population, and leverage randomized assignment to isolate the effect of the treatment.

The simplest mathematical formulation underlying randomized experiment design is as follows. We assume a group of n individuals $i = 1, \ldots, n$. Suppose for an individual, if we apply a treatment, the quantity of interest is equal to a value $Y_1(i)$. If we don't apply the treatment, the quantity of interest is equal to $Y_0(i)$. We define an outcome $Y(i)$ which is equal to $Y_1(i)$ if the treatment is applied and equal to $Y_0(i)$ if the treatment is not applied. In our vaccine example, $Y_1(i)$ indicates whether a person contracts the disease in a specified time period following a vaccination and $Y_0(i)$ indicates whether a person contracts the disease in the same time period absent vaccination. Now, obviously, one person can only take one of these paths. Nonetheless, we can imagine two *potential outcomes*: one

potential outcome $Y_1(i)$ if the treatment is applied and another potential outcome $Y_0(i)$ if the treatment is not applied. Throughout this section, we assume that the potential outcomes are fixed deterministic values.

We can write the relationship between observed outcome and potential outcomes as a mathematical equation by introducing the boolean treatment indicator $T(i)$ which is equal to 1 if subject i receives the treatment and 0 otherwise. In this case, the outcome for individual i equals

$$Y(i) = T(i)Y_1(i) + (1 - T(i))Y_0(i)\,.$$

That is, if the treatment is applied, we observe $Y_1(i)$ and if the treatment is not applied we observe $Y_0(i)$. While this potential outcomes formulation is tautological, it lets us apply the same techniques we use for estimating the mean to the problem of estimating treatment effects.

The individual treatment effect is a relation between the quantities $Y_1(i)$ and $Y_0(i)$, commonly just the difference $Y_1(i) - Y_0(i)$. If the difference is positive, we see that applying the treatment increases the outcome variable for this individual. But, as we've discussed, our main issue is that we can never simultaneously observe $Y_1(i)$ and $Y_0(i)$. Once we choose whether to apply the treatment or not, we can only measure the corresponding treated or untreated condition.

While it may be daunting to predict the treatment effect at the level of each individual, statistical algorithms can be applied to estimate average treatment effects across the population. There are many ways to define a measure of the effect of a treatment on a population. For example, we earlier define the notion of an average treatment effect. Let $\bar{Y}_1$ and $\bar{Y}_0$ denote the means of $Y_1(i)$ and $Y_0(i)$, respectively, averaged over $i = 1, \ldots, n$. We can write

$$\text{Average Treatment Effect} = \bar{Y}_1 - \bar{Y}_0$$

In the vaccine example, this would be the difference in the probability of contracting the illness if one was vaccinated vs if one was not vaccinated.

The *odds* that an individual catches the disease is the number of people who catch the disease divided by the number who do not. The *odds ratio* for a treatment is the odds when every person receives the vaccine divided by the odds when no one receives the vaccine. We can write this out as a formula in terms of our quantities $\bar{Y}_1$ and $\bar{Y}_0$. When the potential outcomes take on values 0 or 1, the average $\bar{Y}_1$ is the number of individuals for which $Y_1(i) = 1$ divided by the total number of individuals. Hence, we can write the odds ratio as

$$\text{Odds Ratio} = \frac{\bar{Y}_1}{1 - \bar{Y}_1} \cdot \frac{1 - \bar{Y}_0}{\bar{Y}_0}\,.$$

This measures the decrease (or increase!) of the odds of a bad event happening when the treatment is applied. When the odds ratio is less than 1, the odds of a bad event are lower if the treatment is applied. When the odds ratio is greater than 1, the odds of a bad event are higher if the treatment is applied.

Similarly, the *risk* that an individual catches the disease is the ratio of the number of people who catch the disease to the total population size. Risk and odds are similar quantities, but some disciplines prefer one to the other by convention. The *risk ratio* is the fraction of bad events when a treatment is applied divided by the fraction of bad events when not applied:

$$\text{Risk Ratio} = \frac{\bar{Y}_1}{\bar{Y}_0}$$

The risk ratio measures the increase or decrease of relative risk of a bad event when the treatment is applied. In the recent context of vaccines, this ratio is popularly reported differently. The *effectiveness* of a treatment is one minus the risk ratio. This is precisely the number used when people say a vaccine is 95% effective. It is equivalent to saying that the proportion of those treated who fell ill was 20 times less than the proportion of those not treated who fell ill. Importantly, it does not mean that one has a 5% chance of contracting the disease.

Estimating treatment effects using randomization

Let's now analyze how to estimate these effects using a randomized procedure. In a randomized controlled trial a group of n subjects is randomly partitioned into a *control group* and a *treatment group*. We assume participants do not know which group they were assigned to and neither do the staff administering the trial. The treatment group receives an actual treatment, such as a drug that is being tested for efficacy, while the control group receives a placebo identical in appearance. An outcome variable is measured for all subjects.

Formally, this means each $T(i)$ is an unbiased coin toss. Because we randomly assign treatments we have

$$\mathbb{E}[Y(i) \mid T(i) = 1] = Y_1(i) \qquad \text{and} \qquad \mathbb{E}[Y(i) \mid T(i) = 0] = Y_0(i)\,.$$

Therefore, for treatment value $t \in \{0, 1\}$,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y(i) \mid T(i) = t\right] = \bar{Y}_t\,.$$

In other words, to get an unbiased estimate of $\bar{Y}_t$ we just have to average out all outcomes for subjects with treatment assignment t. This in turn gives us various causal effects we discussed previously. We can also apply more statistics to this estimate to get confidence bounds and large deviation bounds. Various things we know for estimating the mean of a population carry over. For example, we need the outcome variables to have bounded range in order for our estimates to have low variance. Similarly, if we are trying to detect a tiny causal effect, we must choose n sufficiently large.

Typically in a randomized control trial, the n subjects are supposed to a uniformly random sample from a larger target population of N individuals. The group average $\bar{Y}_t$ is therefore itself only an estimate of the population mean. Here, too, conventional statistics applies in reasoning about how close $\bar{Y}_t$ is to the population average.

Uniform sampling from a population is an idealization that is hard to achieve in experimental practice. It is not only hard to independently sample individuals in a large population, but we also need to be able to set up identical scenarios to test interventions. For medical treatments, what if there is variance between the treatment effect at 9AM in the Mayo Clinic on a Tuesday and at 11PM in the Alta Bates Medical Center on a Saturday? If there are temporal or spatial or other variabilities, the effective size of the population and the corresponding variance grow. Accounting for such variability is a daunting challenge of modern medical and social research that can be at the root of many failures of replication.

The formulation here also assumes that the potential outcomes $Y_t(i)$ do not vary over time. The framework could be generalized to account for temporal variation, but such a generalization will not illuminate the basic issues of statistical methods and modeling. We return to the practice of causal inference and its challenges in the next chapter. But before we do so, we will relate what we just learned to the structural causal models that we saw earlier.

Counterfactuals

Fully specified structural causal models allow us to ask causal questions that are more delicate than the mere effect of an action. Specifically, we can ask *counterfactual* questions such as: Would I have avoided the traffic jam had I taken a different route this morning?

Formally, counterfactuals are random variables that generalize the potential outcome variables we saw previously. Counterfactuals derive from a structural causal model, which gives as another useful way to think about potential outcomes. The procedure for extracting counterfactuals from a

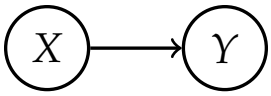

Figure 7: Causal diagram for our traffic scenario.

structural causal model is algorithmic, but it can look a bit subtle at first. It helps to start with a simple example.

A simple counterfactual

Assume every morning we need to decide between two routes $T = 0$ and $T = 1$. On bad traffic days, indicated by $U = 1$, both routes are bad. On good days, indicated by $U = 0$, the traffic on either route is good unless there was an accident on the route.

Let's say that $U \sim B(1/2)$ follows the distribution of an unbiased coin toss. Accidents occur independently on either route with probability $1/2$. So, choose two Bernoulli random variables $U_0, U_1 \sim B(1/2)$ that tell us if there is an accident on route 0 and route 1, respectively. We reject all external route guidance and instead decide on which route to take uniformly at random. That is, $T := U_T \sim B(1/2)$ is also an unbiased coin toss.

Introduce a variable $Y \in \{0, 1\}$ that tells us whether the traffic on the chosen route is good ($Y = 0$) or bad ($Y = 1$). Reflecting our discussion above, we can express Y as

$$Y := T \cdot \max\{U, U_1\} + (1 - T) \max\{U, U_0\} .$$

In words, when $T = 0$ the first term disappears and so traffic is determined by the larger of the two values U and U_0. Similarly, when $T = 1$ traffic is determined by the larger of U and U_1.

Now, suppose one morning we have $T = 1$ and we observe bad traffic $Y = 1$. Would we have been better off taking the alternative route this morning?

A natural attempt to answer this question is to compute the likelihood of $Y = 0$ after the do-operation $T := 0$, that is, $\mathbb{P}_{M[T:=0]}(Y = 0)$. A quick calculation reveals that this probability is $\frac{1}{2} \cdot \frac{1}{2} = 1/4$. Indeed, given the substitution $T := 0$ in our model, for the traffic to be good we need that $\max\{U, U_0\} = 0$. This can only happen when both $U = 0$ (probability $1/2$) and $U_0 = 0$ (probability $1/2$).

But this isn't the correct answer to our question. The reason is that we took route $T = 1$ and observed that $Y = 1$. From this observation, we can deduce that certain background conditions did not manifest for they are inconsistent with the observed outcome. Formally, this means that certain

settings of the noise variables (U, U_0, U_1) are no longer feasible given the observed event $\{Y = 1, T = 1\}$. Specifically, if U and U_1 had both been zero, we would have seen no bad traffic on route $T = 1$, but this is contrary to our observation. In fact, the available evidence $\{Y = 1, T = 1\}$ leaves only the following settings for U and U_1:

Table 3: Possible noise settings after observing evidence. We leave out U_0 from the table, since its distribution is unaffected by our observation.

U	U_1
0	1
1	1
1	0

Each of these three cases is equally likely, which in particular means that the event $U = 1$ now has probability $2/3$. In the absence of any additional evidence, recall, $U = 1$ had probability $1/2$. What this means is that the observed evidence $\{Y = 1, T = 1\}$ has biased the distribution of the noise variable U toward 1. Let's use the letter U' to refer to this biased version of U. Formally, U' is distributed according to the distribution of U conditional on the event $\{Y = 1, T = 1\}$.

Working with this biased noise variable, we can again entertain the effect of the action $T := 0$ on the outcome Y. For $Y = 0$ we need that $\max\{U', U_0\} = 0$. This means that $U' = 0$, an event that now has probability $1/3$, and $U_0 = 0$ (probability $1/2$ as before). Hence, we get the probability $1/6 = 1/2 \cdot 1/3$ for the event that $Y = 0$ under our do-operation $T := 0$, and after updating the noise variables to account for the observation $\{Y = 1, T = 1\}$.

To summarize, incorporating available evidence into our calculation decreased the probability of no traffic ($Y = 0$) when choosing route 0 from $1/4$ to $1/6$. The intuitive reason is that the evidence made it more likely that it was generally a bad traffic day, and even the alternative route would've been clogged. More formally, the event that we observed biases the distribution of exogenous noise variables.

We think of the result we just calculated as the *counterfactual* of choosing the alternative route given the route we chose had bad traffic.

The general recipe

We can generalize our discussion of computing counterfactuals from the previous example to a general procedure. There were three essential steps.

First, we incorporated available observational evidence by biasing the exogenous noise variables through a conditioning operation. Second, we performed a do-operation in the structural causal model after we substituted the biased noise variables. Third, we computed the distribution of a target variable.

These three steps are typically called *abduction*, *action*, and *prediction*, as can be described as follows.

Definition 2. *Given a structural causal model M, an observed event E, an action $T := t$ and target variable Y, we define the* counterfactual $Y_{T:=t}(E)$ *by the following three step procedure:*

1. ***Abduction:*** *Adjust noise variables to be consistent with the observed event. Formally, condition the joint distribution of $U = (U_1, ..., U_d)$ on the event E. This results in a biased distribution U'.*
2. ***Action:*** *Perform do-intervention $T := t$ in the structural causal model M resulting in the model $M' = M[T := t]$.*
3. ***Prediction:*** *Compute target counterfactual $Y_{T:=t}(E)$ by using U' as the random seed in M'.*

It's important to realize that this procedure *defines* what a counterfactual is in a structural causal model. The notation $Y_{T:=t}(E)$ denotes the outcome of the procedure and is part of the definition. We haven't encountered this notation before.

Put in words, we interpret the formal counterfactual $Y_{T:=t}(E)$ as the value Y would've taken had the variable T been set to value t in the circumstances described by the event E.

In general, the counterfactual $Y_{T:=t}(E)$ is a random variable that varies with U'. But counterfactuals can also be deterministic. When the event E narrows down the distribution of U to a single point mass, called *unit*, the variable U' is constant and hence the counterfactual $Y_{T:=t}(E)$ reduces to a single number. In this case, it's common to use the shorthand notation $Y_t(u) = Y_{T:=t}(\{U = u\})$, where we make the variable t implicit, and let u refer to a single unit. The counterfactual random variable Y_t refers to $Y_t(u)$ for a random draw of the noise variables u.

The motivation for the name *unit* derives from the common situation where the structural causal model describes a population of entities that form the atomic units of our study. It's common for a unit to be an individual (or the description of a single individual). However, depending on application, the choice of units can vary. In our traffic example, the noise variables dictate which route we take and what the road conditions are.

Answers to counterfactual questions strongly depend on the specifics of the structural causal model, including the precise model of how the

exogenous noise variables come into play. It's possible to construct two models that have identical graph structures, and behave identically under interventions, yet give different answers to counterfactual queries.[5]

Potential outcomes

Let's return to the *potential outcomes* framework that we introduced when discussing randomized experiments. Rather than deriving potential outcomes from a structural causal model, we assume their existence as ordinary random variables, albeit some unobserved. Specifically, we assume that for every unit u there exist random variables $Y_t(u)$ for every possible value of the assignment t. This potential outcome turns out to equal the corresponding counterfactual derived from the structural equation model:

$$\text{potential outcome } Y_t(u) = Y_{T:=t}(\{u\}) \text{ structural counterfactual}$$

In particular, there is no harm in using our potential outcome notation $Y_t(u)$ as a shorthand for the corresponding counterfactual notation.

In the potential outcomes model, it's customary to think of a binary *treatment variable* T that assumes only two values, 0 for *untreated*, and 1 for *treated*. This gives us two potential outcome variables $Y_0(u)$ and $Y_1(u)$ for each unit u. There is some potential for notational confusion here. Readers already familiar with the potential outcomes model may be used to the notation "$Y_i(0), Y_i(1)$" for the two potential outcomes corresponding to unit i. In our notation the unit (or, more generally, set of units) appears in the parentheses and the subscript denotes the substituted value for the variable we intervene on.

The key point about the potential outcomes model is that we only observe the potential outcome $Y_1(u)$ for units that were treated. For untreated units we observe $Y_0(u)$. In other words, we can never simultaneously observe both, although they're both assumed to exist in a formal sense. Formally, the outcome $Y(u)$ for unit u that we observe depends on the binary treatment $T(u)$ and is given by the expression:

$$Y(u) = Y_0(u) \cdot (1 - T(u)) + Y_1(u) \cdot T(u)$$

We can revisit our traffic example in this framework. The next table summarizes what information is observable in the potential outcomes model. We think of the route we choose as the treatment variable, and the observed traffic as reflecting one of the two potential outcomes.

Table 4: Traffic example in the potential outcomes model

Route T	Outcome Y_0	Outcome Y_1	Probability
0	0	?	1/8
0	1	?	3/8
1	?	0	1/8
1	?	1	3/8

Often this information comes in the form of samples. For example, we might observe the traffic on different days. With sufficiently many samples, we can estimate the above frequencies with arbitrary accuracy.

Table 5: Traffic data in the potential outcomes model

Day	Route T	Outcome Y_0	Outcome Y_1
1	0	1	?
2	0	0	?
3	1	?	1
4	0	1	?
5	1	?	0
...	...	...	...

In our original traffic example, there were 16 units corresponding to the background conditions given by the four binary variables U, U_0, U_1, U_T. When the units in the potential outcome model agree with those of a structural causal model, then causal effects computed in the potential outcomes model agree with those computed in the structural equation model. The two formal frameworks are perfectly consistent with each other.

As is intuitive from the table above, causal inference in the potential outcomes framework can be thought of as filling in the missing entries ("?") in the table above. This is sometimes called *missing data imputation* and there are numerous statistical methods for this task. If we could *reveal* what's behind the question marks, many quantities would be readily computable. For instance, estimating the average treatment effect would be as easy as counting rows.

When we were able to directly randomize the treatment variable, we showed that treatment effects could be imputed from samples. When we are working with observational data, there is a set of established conditions under which causal inference becomes possible:

1. **Stable Unit Treatment Value Assumption** (SUTVA): The treatment

that one unit receives does not change the effect of treatment for any other unit.

2. **Consistency**: Formally, $Y(u) = Y_0(u)(1 - T(u)) + Y_1(u)T(u)$. That is, $Y(u) = Y_0(u)$ if $T(u) = 0$ and $Y(u) = Y_1(u)$ if $T(u) = 1$. In words, the outcome $Y(u)$ agrees with the potential outcome corresponding to the treatment indicator.
3. **Ignorability**: The potential outcomes are independent of treatment given some deconfounding variables Z, i.e., $T \perp (Y_0, Y_1) \mid Z$. In words, the potential outcomes are conditionally independent of treatment given some set of deconfounding variables.

The first two assumptions automatically hold for counterfactual variables derived from structural causal models according to the procedure described above. This assumes that the units in the potential outcomes framework correspond to the atomic values of the background variables in the structural causal model.

The third assumption is a major one. The assumption on its own cannot be verified or falsified, since we never have access to samples with both potential outcomes manifested. However, we can verify if the assumption is consistent with a given structural causal model, for example, by checking if the set Z blocks all backdoor paths from treatment T to outcome Y.

There's no tension between structural causal models and potential outcomes and there's no harm in having familiarity with both. It nonetheless makes sense to say a few words about the differences of the two approaches.

We can derive potential outcomes from a structural causal model as we did above, but we cannot derive a structural causal model from potential outcomes alone. A structural causal model in general encodes more assumptions about the relationships of the variables. This has several consequences. On the one hand, a structural causal model gives us a broader set of formal concepts (causal graphs, mediating paths, counterfactuals for every variable, and so on). On the other hand, coming up with a plausibly valid structural causal model is often a daunting task that might require knowledge that is simply not available. Difficulty to come up with a plausible causal model often exposes unsettled substantive questions that require resolution first.

The potential outcomes model, in contrast, is generally easier to apply. There's a broad set of statistical estimators of causal effects that can be readily applied to observational data. But the ease of application can also lead to abuse. The assumptions underpinning the validity of such estimators are experimentally unverifiable. Our next chapter dives deeper into the practice of causal inference and some of its limitations.

Chapter notes

This chapter was developed and first published by Barocas, Hardt, and Narayanan in the textbook *Fairness and Machine Learning: Limitations and Opportunities*.[6] With permission from the authors, we include a large part of the original text here with only slight modifications. We removed a significant amount of material on discrimination and fairness and added an extended discussion on randomized experiments.

There are several excellent introductory textbooks on the topic of causality. For an introduction to causality with an emphasis on causal graphs and structural equation models turn to Pearl's primer,[7] or the more comprehensive textbook.[4] Our exposition of Simpson's paradox and the UC Berkeley data was influenced by Pearl's discussion, updated for a new popular audience book.[8] The example has been heavily discussed in various other writings, such as Pearl's recent discussion.[8] We retrieved the Berkeley data from `http://www.randomservices.org/random/data/Berkeley.html`. There is some discrepancy with the data available on the Wikipedia page for Simpson's paradox that we retrieved on Dec 27, 2018.

For further discussion regarding the popular interpretation of Simpson's original article,[9] see the article by Hernán, Clayton, and Keiding,[10] as well as Pearl's text.[4]

The technically-minded reader will enjoy complementing Pearl's book with the recent open access text by Peters, Janzing, and Schölkopf[5] that is available online. The text emphasizes two variable causal models and applications to machine learning. See Spirtes, Glymour and Scheines[11] for a general introduction based on causal graphs with an emphasis on *graph discovery*, i.e., inferring causal graphs from observational data. An article by Schölkopf provides additional context about the development of causality in machine learning.[12]

The classic formulation of randomized experiment design due to Jerzy Neyman is now subsumed by and commonly referred to as the framework of potential outcomes.[13,14] Imbens and Rubin[15] give a comprehensive introduction to the technical repertoire of causal inference in the potential outcomes model. Angrist and Pischke[16] focus on causal inference and potential outcomes in econometrics. Hernán and Robins[17] give another detailed introduction to causal inference that draws on the authors' experience in epidemiology. Morgan and Winship[18] focus on applications in the social sciences.

Bibliography

[1] Peter J. Bickel, Eugene A. Hammel, and J. William O'Connell. Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404, 1975.

[2] Linda L. Humphrey, Benjamin K. S. Chan, and Harold C. Sox. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. *Annals of Internal Medicine*, 137(4):273–284, 08 2002.

[3] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *International Journal of Epidemiology*, 43(2):511–515, 2014. Reprint.

[4] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[5] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

[7] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

[8] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[9] Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

[10] Miguel A. Hernán, David Clayton, and Niels Keiding. The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3):780–785, 03 2011.

[11] Peter Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search*. MIT Press, 2000.

[12] Bernhard Schölkopf. Causality for machine learning. *arXiv:1911.10500*, 2019.

[13] Jerzy Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[14] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[15] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

[16] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

[17] Miguel A. Hernán and James Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

[18] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2014.