DS 102 Data, Inference, and Decisions **Spring 2020**

Lecture 13: Causal Inference I

Lecturer: Moritz Hardt

13.1 Why Causal Inference?

Today, we will begin our discussion of causal inference, which gives us a formal way to think about cause and effect. Oftentimes when we perform statistical inferences, we talk about correlation between random variables rather than *causation*. Depending on our application, correlation may not be enough. But, what is causation? In the next sections, we will build up a framework to formalize what we mean by "causation" or a "causal effect."

13.2 A Motivating Example

We will begin with a concrete example that highlights differences between correlation and causation.

Example 13.1. Suppose we are given the following data from a local hospital about the success rates of two kidney stone treatments:

	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Combined	78% (273/350)	83% (289/350)

We can make the following observations:

- Within the subgroup of patients that had small kidney stones, Treatment A has a higher success rate.
- Within the subgroup of patients that has large kidney stones, Treatment A has a higher success rate.
- Treatment B has a higher success rate across the entire patient population.

These observations appear to be contradictory. How is it possible that Treatment B is more effective across the whole population, but less effective for both subgroups?

We cannot know from this data alone which treatment, A or B, is truly better. In order to make such a judgement, we need to think more deeply about what is going on, how the data were

generated, and what potential confounders exist. In particular, we need to understand the "causal story" behind the data.

Here, is one potential causal story behind the data in our table: Perhaps Treatment B is less costly and less invasive (e.g. an oral medication), so it is more commonly used for less severe cases, including most cases of small kidney stones. Doctors thus prefer to avoid Treatment A—which is highly effective but much more costly and invasive (e.g. a surgery)—in all but the most difficult cases.

This causal story explains the data well, but the data alone do not tell us about cause and effect. More information (and some careful thought about potential underlying factors) is needed to come up with a causal story.

Example 13.1 is an example of Simpson's paradox. Formally, if Y is the outcome and X is a binary variable indicating subgroup membership, we have

$$\begin{aligned}&\mathbb{P}(Y|A) < \mathbb{P}(Y|B),\\ &\mathbb{P}(Y|A,X) > \mathbb{P}(Y|B,X), \text{ and}\\ &\mathbb{P}(Y|A,\neg X) > \mathbb{P}(Y|B,\neg X)\end{aligned}$$

all hold, where $\neg X$ means "not X." Mathematically, there is no contradiction here, and yet it conflict with our intuition. In particular, this tension arises when we misinterpret the conditional probabilities in our data as actions. When we see "Treatment A," we tend to think that means we *administered* Treatment A, but in reality it means we *observe* a doctor giving Treatment A. Conditional probability should be interpreted as instructions to do inference—in Example 13.1, if we observe that you got Treatment A, we can infer you likely had a large kidney stone, but we do not make any intervention or take any action.

In Example 13.1, our data comes from observing doctors in a hospital, which means that treatments are assigned according to doctors' inclinations and the current standards of care. Doctors assign treatment based in part on the severity of a patient's case, but the severity of a patient's case *also* influences their chances of successful recovery. We can contrast this with a **randomized trial**, wherein we randomly assign treatments to patients regardless of kidney stone size, thereby breaking the natural practice of the doctor. Since we toss a coin to determine treatment, we break whatever relationship there might be between treatment assignment and severity.

In general, in a randomized trial we take an *action* and make an *active assignment*. The effects of this action are in general not given by conditional probability—conditioning on something is not the same as performing an action in the real world. We call the effect of an action its **causal effect**.

In the remaining sections, we will focus on building up a new language of causation—one which is different from the language of probability. This new language will allow us to formalize what it means to take actions, and for an action to have a causal effect.

13.3 Intuition from Programming

Suppose we have a program to generate a distribution step by step. We can think of this like a Python file or Jupyter notebook that we can run to generate data. Intuitively, we would rather have a program for data generation than just a dataset; we can actually see how data were generated, and see how changing lines of code affects the outcomes we're interested in. Moreover, it is easy to go from the generating program to a dataset, but not the other way around.

Example 13.2. Suppose we have the following data generating program:

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

We might choose to interpret X as whether or not an individual exercises, W as whether or not they are overweight, and H as whether or not they have heart disease. This program defines a joint distribution over X, W, H, and we can compute probabilities in this joint distribution. For example,

$$\mathbb{P}(H = 1) = \mathbb{P}(\neg X, \neg U_3) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

and

$$\mathbb{P}(H = 1 | W = 1) = \frac{1}{3}.$$

The conditional probability of heart disease ($H = 1$) given being overweight ($W = 1$) is higher than the overall probability of heart disease. Does this mean that being overweight causes heart disease in this model? We know that this is not the case because in our model H is determined by X and U_3, not W. We see that conditional probabilities are misleading, and not quite the right way to capture whether or not H is caused by W.

What should we do instead? One thing we can do to really drive home the fact that H does not depend on W is to actively set W to some value and see what happens. For instance, consider substituting $W = 1$:

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

In this new program, the probability of $H = 1$ is still $\frac{1}{6}$, so the assignment we made for W did not affect the probability of H at all. We cannot express the probability of H after making an assignment using typical conditional probability, so we need to introduce new notation:

$$\mathbb{P}(H = 1|\mathbf{do}(W = 1)) = \frac{1}{6}.$$

This is called **do-substitution** or **do-intervention**, and the **do** is called the **do-operator**.

The most important take-away from Example 13.2 is that, in general, the probability of H conditional on observing $W = 1$ is *not* the same as the probability of H when we intervene to *set* $W = 1$:

$$\mathbb{P}(H|W = 1) \neq \mathbb{P}(H|\mathbf{do}(W = 1)).$$

Observing is not the same as *doing*. In the next lecture, we will see some ways of actually estimating probabilities of the form $\mathbb{P}(H|\mathbf{do}(W = 1))$.

13.4 Structural Causal Models and Causal Graphs

The "programs" we saw in the previous section are known formally as **structural causal models**. Each structural causal model can be associated with an acyclic assignment graph, called a **causal graph**, which gives the dependence structure in the model. For example, Figure 13.1 gives the causal graph for our simple model in Example 13.2. We interpret directed paths in the causal graph as causal links, which means that all the causes of a given variable are all of its ancestors in the causal graph.

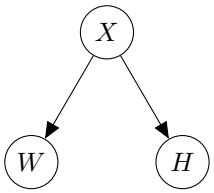

Figure 13.1: Causal graph for the structural causal model in Example 13.2, which shows that X has a causal effect on both W and H. However, W does not have a causal effect on H since there is no directed path $W \to H$.

13.5 Confounding

We say that the variables X and Y are confounded if

$$\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|\mathbf{do}(X = x)),$$

and people use the phrase **confounding bias** to express the fact that this has occurred. In terms of the causal graph, confounding corresponds to a graph structure where a confounding variable Z has an influence on both the explanatory variable X and the response Y, as in Figure 13.2.

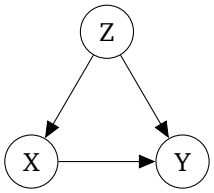

Figure 13.2: Causal graph where Z is a confounding variable.

We have already seen an example which corresponds to such a causal graph. In Example 13.1, based on our causal story, a patient's treatment and their successful recovery were confounded, since kidney stone size influenced both which treatment was administered and chances of success. This is captured by the causal graph in Figure 13.3, which has the same form as Figure 13.2.

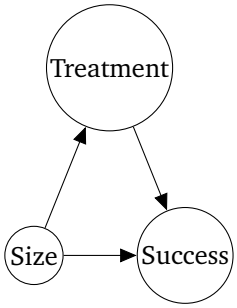

Figure 13.3: Causal graph for our motivating kidney stone example (Example 13.1).

How can we eliminate the confounding influence of Z? One thing we could do is to assign treatment X randomly, as in a randomized trial, which would destroy the link between the confounder Z and the treatment X. More generally, to eliminate confounding, we need to hold the confounding variable constant in our analyses. For instance, in our kidney stone example, we should hold the size of the stone constant while performing inferences. In general, the right thing to do is to "slice" our data by the values of the variable Z, which is called "controlling for the variable Z."

One important caveat: we must be careful not to control for **mediators**. Figure 13.4 gives a causal graph where the variable Z is a mediating variable. In this case, there is a direct path $X \to Y$ and an indirect path $X \to Z \to Y$. The path $X \to Z \to Y$ is a legitimate causal pathway from X to Y, so in this case blocking the path by holding Z constant reduces the actual causal effect of X on Y.

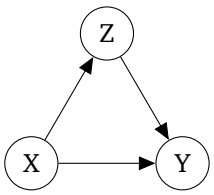

Figure 13.4: Causal graph where Z is a mediating variable.

13.6 The Big Picture

In general, data alone does not tell us the full story; just looking at a bunch of numbers cannot tell us about causal structure, and thus cannot help us draw conclusions about causal effects.

To get at causal effects, we need to make assumptions about the causal story. We represent these causal stories formally using structural causal models and/or causal graphs. These causal models and causal graphs represent assumptions that we make about the world; they are not given to us, nor is there any way to get them out of our data. Instead, they come from assumptions that we make based on domain knowledge. For instance, in Example 13.1, we might choose our causal graph by talking to doctors about their decision-making processes when they assign treatments to kidney stone patients.

Once we have the causal graph as a representation of the causal story, we can use it to decide which variables are mediators, which variables are confounders, and thus which variables we should control for. We should try to control for confounding variables, but not for mediating variables. The idea is that, once we have a causal model, we can leverage it interpret our data correctly.

In the next lecture, we will see how causal inference plays out in practice, and some techniques for actually estimating causal effects from observational data.