36-705: Intermediate Statistics **Fall 2019**

Lecture 28: November 8

Lecturer: Siva Balakrishnan

In this and the next lecture we will try to make sense of some basic questions in causal inference.

28.1 Causal Inference

A lot of statistics focusses on questions of association. Are X and Y correlated? Is X predictive of Y, and so on.

In many applications however, our questions are inherently causal: in medicine we wish to know if a new drug is effective against a disease. This is not a question of association, because if I went out in the world and measured all the people taking aspirin, most likely many of them would have headaches so I could (correctly?) conclude aspirin and headaches are associated. It is almost certainly not the case that aspirin causes headaches and this is what we usually mean by the phrase: "correlation does not imply causation."

It will take a bit of work to get to the questions of interest but broadly you should think of the two statistical questions in causal inference as analogous to ones we have considered so far: we want to estimate the causal effect (point estimation) and construct confidence intervals for the causal effect (inference/hypothesis testing).

28.2 The Potential Outcomes Framework

The basic language of causal inference that we will adopt comes from the work of Neyman (and later Rubin). Causality is tied to something known as a manipulation/intervention applied to a *unit* (think person).

We will think of the case when there are two possible actions (or treatments). Think of taking an aspirin and not taking an aspirin as the two treatments. Often we refer to one of the treatments as the active treatment (or just treatment) and the other as the control treatment (or just control).

We associate every unit and the two treatments with two *potential outcomes*: the potential outcome if the unit received the treatment and the potential outcome if the unit received control. A priori both potential outcomes are possible. However, every unit only receives one of the two treatments (i.e. either treatment or control) and so we only observe one of

the two potential outcomes. This is known as the fundamental problem of causal inference. We only observe one of the potential outcomes for each unit.

While all of this might seem rather obvious, thinking formally about treatment and control, and the potential outcomes is extremely important to causal inference. A point of particular emphasis is that if you are asking a causal question, ideally you need to be able to meaningfully say what the "treatment" is and what the potential outcomes are.

Here are a few examples of statements:

1. "Aspirin cures headaches." In order to cast this is the potential outcomes framework we could imagine that for a person with a headache (a unit) we could either give the person aspirin (treatment) or a placebo (control), and observe the corresponding potential outcome.

2. "She has long hair because she is a girl." This sounds like a causal statement so we should be able to describe the experiment. Is a unit a girl/boy? What exactly is a treatment? Can we meaningfully say what the potential outcomes are?

For some causal questions we can naturally define an associated "experiment". Murky causal questions are ubiquitous, and are in some sense interesting and challenging. For instance, I might like to know the effect of race on life expectancy. If you go through the exercise above again you will have a lot of trouble. Research in social science, political science, epidemiology and economics (to name a few fields) are centered around how to make sense of these difficult questions.

We will focus on simpler case, where there are well-defined interventions and potential outcomes.

In this case, for the i^{th} unit we will denote the potential outcome if the unit receives control as $Y_i(0)$ and the potential outcome if the unit receives treatment as $Y_i(1)$. A natural definition of the *causal effect* of treatment on the i^{th} unit is $Y_i(1) - Y_i(0)$ (you could consider any other meaningful function of the potential outcomes and we will discuss this soon). Again, the fundamental problem of causal inference is that we only observe $Y_i(1)$ or $Y_i(0)$ and not both.

28.3 Causal Estimands

Finally, let us be a bit more precise about what we'd like to estimate. There are many things we might care about estimating:

1. Unit level causal effects: things like $Y_i(1) - Y_i(0)$ or $Y_i(1)/Y_i(0)$.

2. The average treatment effect:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)).$$

   This is what we will focus on in this class.

3. Average treatment effect over sub-populations:

$$\tau_S = \frac{1}{|S|}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))\mathbb{I}(i \in S).$$

   For instance the set S could be all men in the population (i.e. I am interested in whether aspiring relieves headaches in men).

28.4 The Average Treatment Effect

In the simplest setting we are interested in understanding if there is a causal link between a binary treatment W and an outcome Y. For a particular unit, if $W_i = 1$ we say that the unit is treated and if $W_i = 0$ then the unit is in the control group.

There are many ways to measure causal associations. The one we will adopt is to say there are two potential outcomes $Y(0)$ and $Y(1)$ which are respectively the outcomes if a unit is in the control group and if it is treated. We will then measure the casual association by some function of the potential outcomes: a typical one is the average treatment effect, i.e.

$$\tau = \mathbb{E}[Y(1) - Y(0)],$$

which is the difference in outcomes if all units were treated versus all were in the control group. If we are dealing with a finite population we might instead define:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - Y_i(0)].$$

The main problem in causal inference is that each unit is either treated or in the control group so we never observe both potential outcomes. What we do observe is:

$$Y_i^{\text{obs}} = Y_i(1)W_i + Y_i(0)(1 - W_i).$$

Suppose that m individuals are treated. A quantity we can estimate is:

$$\alpha = \mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = 0]$$

where we use the estimator

$$\widehat{\alpha} = \frac{1}{m} \sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{n-m} \sum_{i:W_i=0} Y_i^{\text{obs}}, \tag{28.1}$$

since this only depends on the observed data. In general, $\alpha \neq \tau$ since in a typical setting we have selection bias, i.e. people can choose treatment or control based on their knowledge of their potential outcomes so that W and $(Y(0), Y(1))$ are not independent. One formal way of defining selection bias in this context is simply as the difference between τ and α. A basic question is then: when are α and τ the same?

Correlation is not causation: You often hear this phrase. Perhaps, one way to think about this statement is that it is simply saying $\alpha \neq \tau$. Suppose we temporarily defined $W_i = -1$ if a unit is in control (this does not change anything), then,

$$\alpha = \mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = -1],$$

is roughly the *correlation* between the observed outcomes and the treatment W (strictly speaking, it is the correlation when the groups are balanced). Concretely,

$$\mathbb{E}[Y^{\text{obs}}W] = \mathbb{E}[Y(1)|W = 1]P(W = 1) - \mathbb{E}[Y(0)|W = -1]P(W = -1).$$

If we can ensure that

$$W \perp\!\!\!\perp (Y(0), Y(1))$$

then we indeed have that,

$$\begin{aligned} \alpha &= \mathbb{E}[Y(1)|W = 1] - \mathbb{E}[Y(0)|W = 0] \\ &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \tau. \end{aligned}$$

The general way to ensure this is via a randomized trial. In this case, we *randomly* assign people to take the treatment or control and this ensures the above independence.

We will consider the case for the next couple of sections of a completely randomized trial, i.e. we select m individuals uniformly at random assign them treatment and the remaining are assigned control. We use the estimator in (28.1).

Now, we can see that this is an unbiased estimator of the average treatment effect. It is worth noting that the only thing that is surely random here is our treatment assignment, the potential outcomes can be fixed or random (it does not matter which).

$$\mathbb{E}[\widehat{\tau}] = \sum_{i=1}^{n} \frac{\mathbb{E}\left(W_i\right)}{m} Y_i(1) - \frac{\mathbb{E}\left((1 - W_i)\right)}{n-m} Y_i(0).$$

The mean of W_i is given by:

$$\mathbb{E}(W_i) = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n},$$

and

$$\mathbb{E}(1 - W_i) = \frac{n-m}{n}.$$

This gives us that:

$$\mathbb{E}[\widehat{\tau}] = \tau.$$

In general, constructing p-values or confidence intervals for τ can be challenging. We will consider one special case when we can use a permutation test to construct a p-value.

28.5 Hypothesis testing: Fisher's Exact p-values

Fisher was one of the first statisticians to understand the power of a randomized trial. In agricultural experiments, he advocated randomized experiments in order to draw rigorous causal conclusions.

A natural subsequent problem is: given an estimate of the causal effect, assess its significance (or construct confidence intervals for it).

Fisher gave a way to construct valid p-values under what is called the *sharp null*, i.e. the null hypothesis that for every unit i the potential outcomes are the same under the treatment and control, i.e. the treatment has no effect. The method is reminiscent of the permutation method we used for two-sample testing.

Suppose for simplicity that we are using the estimator described in the previous section and we reject the null hypothesis if $|\widehat{\tau}|$ is large. Under the null hypothesis, we can determine both potential outcomes $Y_i(0)$ and $Y_i(1)$ for all the units.

We can now use the permutation method, suppose a different set T' of m units were to receive treatment: then our estimate would be:

$$\widehat{\tau}_{T'} = \frac{1}{m}\sum_{i \in T'} Y_i(1) - \frac{1}{n-m}\sum_{i \notin T'} Y_i(0),$$

where we can use the sharp null hypothesis to "fill in" the potential outcomes we do not observe. We can repeat this many times (say B) and compute the p-value:

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(|\widehat{\tau}_{T_b}| \geq |\widehat{\tau}|).$$

It is easy to verify that this is a valid p-value.

The intuition is identical to the permutation test, if there was in fact a different in outcomes under treatment and controls (say treatment potential outcomes were much higher than control potential outcomes) then we would expect the p-value to be small, since the difference in means will get smaller when we randomly swap some of the treatment and control outcomes.

28.6 Confounding

For most interesting policy questions, we cannot actually do a randomized trial. For instance, if I wanted to know if smoking caused lung cancer, there are ethical issues with trying to run a randomized trial. In this case, we have to use what is called *observational data*, i.e. we have information about many people who are smokers and not, and whether they have lung cancer or not.

It is clear that we can measure the correlation between smoking and lung cancer: the main question is when, if ever, can we claim a causal relationship?

The main problem is again selection bias.

Here is a motivating example: Suppose that our population has two kinds of people, those who are always healthy ($Y_i(1) = Y_i(0) = 1$) irrespective of whether they take the treatment or not, and those who are always unhealthy ($Y_i(1) = Y_i(0) = 0$) irrespective of whether they take the treatment or not. Suppose further that mostly healthy people take the treatment, while the unhealthy ones do not take the treatment.

The causal effect $\tau = 0$, but the estimator above would yield, $\widehat{\tau} \approx 1$, and we might incorrectly conclude that the treatment is beneficial.

Suppose however, that we knew who the healthy people were and who the unhealthy people were (we could gather such information by asking people questions about their lifestyle and other things). Then we could try to compare healthy people who took the treatment with healthy people who did not and similarly compare unhealthy people who took the treatment with unhealthy people who did not (and then try to combine these two estimates in some way). In this case, when we compared two healthy people who took the treatment and who did not we would see the treatment had no effect, and similarly for the unhealthy ones. We would likely correctly conclude that the treatment has no effect.

The key assumption that makes causal inference from observational data possible is the assumption of *no unmeasured confounding* or *selection on observables* or *ignorability*. Formally, we suppose that we have access to covariates X (think demographic information) such that,

$$W \perp\!\!\!\perp (Y(1), Y(0))|X.$$

This is an assumption. Roughly the assumption is is plausible in settings where we believe

we can measure all of the covariates that explain the decision to take the treatment. We also need some other assumptions (SUTVA from the previous lecture, and the assumption that $\mathbb{P}(W = 1|X = x)$ is bounded away from 0 and 1, i.e. every individual has some non-zero chance of being either treated or in the control group) but we will ignore these and focus on no unmeasured confounding.

One way to think about this assumption, is that conditional on X we have a randomized trial, i.e. the treatment is independent of the potential outcomes. So if we condition on X (they are the confounders) we no longer have any selection bias. Alternatively, within levels of the covariate treatment is decided by (a biased) coin flip.

The other way to think about it is to think about a procedure called *matching*, i.e. suppose temporarily that X is discrete as it was in the healthy/unhealthy example. Then we could *match* people who had the same covariates but different treatments (i.e. some of them received treatment and some of them received control), and take the difference in their outcomes. For these matched people, under the assumption of no unmeasured confounding we are effectively observing *both potential outcomes*.

Setup and Notational Comments: In what follows we will assume we observe triplets from some distribution $(X, W, Y^{\text{obs}}) \sim F$, where as usual

$$Y^{\text{obs}} = WY(1) + (1 - W)Y(0).$$

We will assume all of these are stochastic. Sometimes we will use the subscript $\mathbb{E}_X$ to denote taking an expectation with respect to the randomness of X.

28.7 Identification under no unmeasured confounding

We want to estimate:

$$\tau = \mathbb{E}[Y(1) - Y(0)],$$

and to do this (as we did previously) we need to be able to write it in terms of observable quantities. Notice that by the law of total expectation,

$$\tau = \mathbb{E}_X[\mathbb{E}[Y(1) - Y(0)|X]].$$

Furthermore,

$$\tau = \mathbb{E}_X[\mathbb{E}[Y(1)|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y(0)|X, W = 0]],$$

since we have that $W \perp\!\!\!\perp (Y(1), Y(0))|X$. This can we written as:

$$\tau = \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 0]],$$

which is just a function of the observed data, and so we can turn our attention to estimating this quantity.

28.8 Estimation under no unmeasured confounding

The most direct way to estimate τ is to estimate:

$$\mu_0(x) = \mathbb{E}[Y^{\text{obs}}|X = x, W = 0]$$
$$\mu_1(x) = \mathbb{E}[Y^{\text{obs}}|X = x, W = 1].$$

These are two functions of the covariates X, one of them is the average outcome of the treatment group as a function of the covariates, and the other is the average outcome of the control group as a function of the covariates.

Estimating a conditional expectation is a problem is probably the most common problem in statistics – it is known as *regression*. We will delve into this formally in the next few lectures but for now let us suppose that someone hands us estimators $\widehat{\mu}_0$ and $\widehat{\mu}_1$ of these two functions.

Then we can compute the plug-in estimator:

$$\begin{aligned}\widehat{\tau} &= \widehat{\mathbb{E}}_X \left[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)\right], \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)\right]\end{aligned}$$

which is just the average of the difference between two regression functions. One approximately correct way to think about this is that we are using regression to impute the missing potential outcomes for each individual.

There are other ways to try to estimate τ. The other popular estimator is called the Horvitz-Thompson estimator or the inverse propensity score estimator. I will explain the basic idea here.

First let us define a quantity, known as the *propensity score*:

$$\pi(x) = \mathbb{P}(W = 1|X = x),$$

which represents the probability that a unit with covariates x receives treatment. Note that,

$$\mathbb{E}[W|X = x] = \pi(x)$$
$$\mathbb{E}[1 - W|X = x] = 1 - \pi(x).$$

Returning to the average treatment effect,

$$\begin{aligned}\tau &= \mathbb{E}[Y(1) - Y(0)] \\ &= \mathbb{E}_X[\mathbb{E}[Y(1) - Y(0)|X = x]] \\ &= \mathbb{E}_X\left(\mathbb{E}_W\left[\mathbb{E}\left[\frac{Y(1)W}{\pi(x)}|X = x\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[\frac{Y(0)(1-W)}{1-\pi(x)}|X = x\right]\right]\right).\end{aligned}$$

This in turn can be written in terms of only observed quantities:

$$\begin{aligned}\tau &= \mathbb{E}_X\left(\mathbb{E}_W\left[\mathbb{E}\left[\frac{Y^{\text{obs}}W}{\pi(x)}|X=x\right]\right] - \mathbb{E}_W\left[\mathbb{E}\left[\frac{Y^{\text{obs}}(1-W)}{1-\pi(x)}|X=x\right]\right]\right) \\ &= \mathbb{E}\left[\frac{Y^{\text{obs}}W}{\pi(x)}\right] - \mathbb{E}\left[\frac{Y^{\text{obs}}(1-W)}{(1-\pi(x))}\right]\end{aligned}$$

which we can estimate as:

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i^{\text{obs}}W_i}{\pi(X_i)} - \frac{Y_i^{\text{obs}}(1-W_i)}{1-\pi(X_i)}\right].$$

This is called the Horvitz-Thompson estimator or the inverse propensity score estimator. In practice, we do not know the propensity score so we would estimate it from the data: this is again a problem of regression except the outcome is binary (i.e. since $\pi(X)$ is simply a conditional expectation we can estimate it by regressing W on X).

28.9 Advanced topics

This is just the tip of the iceberg. If you take a course in Causal Inference (next semester!) you will see many other neat things (here is just a small subset):

1. No unmeasured confounding is just one assumption that leads to identification of a causal effect. More broadly, in economics, political science and other fields people look for what are called natural experiments, i.e. roughly some subset of the population for which the assignment to treatment/control is nearly random.

2. Even in a randomized trial you might have something called non-compliance, i.e. some people don't do what they are told. In this case, you need to adjust your estimates. This is a canonical example of something called an instrumental variable problem.

3. There are many things beyond the average treatment effect that you might want to estimate. They all have different assumptions under which they are identified (i.e. can be written in terms of observable quantities) and there are different strategies to estimate them.

4. There is a very nice/simple way to combine the regression-based and propensity-score based estimators from above to construct what are called *doubly robust* estimators. These have the property that they are consistent if you can estimate either the regression function or the propensity score well (i.e. you do not need to estimate both well).

5. From a purely statistical standpoint, the average treatment effect under no unmeasured confounding:

$$\tau = \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 1]] - \mathbb{E}_X[\mathbb{E}[Y^{\text{obs}}|X, W = 0]],$$

   is an example of a functional. The regression method we used is basically the plug-in estimator of the functional. Plug-in estimators for functionals are usually not optimal. Optimally estimating functionals is a challenging statistical problem and there is an analog of the Cramer-Rao, efficiency, score functions, Fisher information etc. for functional estimation problems. This is known as semi-parametric efficiency theory.

6. There are many different languages for talking about causality and causal inference. We used potential outcomes. Many people (including many people in Philosophy here) use a language pioneered by Judea Pearl based on (causal) directed graphs. Most things can be translated from one language to another. People who work on causal graphs often focus effort on a much harder problem of *causal discovery*. We considered the case where there was a treatment and an outcome and we wanted to know if the treatment had a causal effect on the outcome. In causal discovery we have a collection of variables and want to discover causal relationships between them. This requires even stronger assumptions.