36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 28: November 7

Lecturer: Siva Balakrishnan

28.1 Review and Outline

In the last class we discussed independence testing and measuring association:

1. χ^2 test for independence of discrete random variables
2. Measuring association for continuous RVs, Pearson correlation, Spearman's rho, and Mutual Information
3. Independence between one discrete and one continuous RV

Today we will begin our discussion of Causal Inference. This is one of my recent favorite topics in statistics. Some of the material we will cover is in Chapter 16 of the Wasserman book, but I recommend reading the first few chapters of the book "Causal Inference for Statistics, Social and Biomedical Sciences" by Imbens and Rubin.

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). The key idea is that 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?

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 Multiple Units

Defining the causal effect does not require multiple units, however, estimating causal effects does. The idea is simple, suppose I observe the potential outcomes under treatment for some units and the potential outcomes under control for some units, then maybe in some cases I can put these together to get a sense of the average causal effect.

This actually requires another assumption. This is called the Stable-Unit-Treatment-Value-Assumption (SUTVA). The assumption has two parts:

1. Giving treatment/control to one unit does not affect the potential outcomes of other units,
2. For units receiving treatment (or control) there is only one level of treatment (it cannot be that some units take one aspirin, some take two and a few take 10000).

28.4 The assignment mechanism

The next part of the story, is what is called the assignment mechanism. Suppose we have n units $\{1, \ldots, n\}$. The assignment mechanism, is what determines which potential outcome we observe for each unit.

We will denote the assignment vector: $W \in \{0,1\}^n$ where $W_i = 0$ means the unit i is assigned to control and $W_i = 1$ means that unit i is assigned to treatment.

The treatment mechanism that we will focus on today is what is known as a completely randomized trial. This means that we select a number m of units to receive treatment, and then select m out of the n units uniformly at random. In mathematical notation, this means that:

$$\mathbb{P}(W = w) = \frac{1}{\binom{n}{m}},$$

for any binary vector w, with $\sum_i w_i = m$.

An alternative that is popular is something called a Bernoulli trial, where each individual has some fixed probability p (think 0.5) of receiving treatment. Alternatively, you could imagine *stratified* randomized assignments where the units are grouped and then randomly assigned treatment/control within the group. With this notation we can now write the observed and missing potential outcomes:

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

More generally, you could imagine situations where a doctor measures some covariates of the patient (her blood pressure, age, and height say) and then decides whether to recommend the treatment or not. In this case, the assignment is not random, and we will discuss situations like this in a future lecture.

28.5 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).

28.6 Estimating the average treatment effect

Under the assumption of a completely randomized trial (and SUTVA), it is easy to construct an estimator of the average treatment effect. Let us denote the set of treated units as T and the number of treated units as m (remember that m is fixed). Our estimate is:

$$\begin{aligned}\widehat{\tau} &= \frac{1}{m}\sum_{i\in T} Y_i(1) - \frac{1}{n-m}\sum_{i\notin T} Y_i(0) \\ &= \sum_{i=1}^{n}\left(\frac{W_i}{m}Y_i(1) - \frac{(1-W_i)}{n-m}Y_i(0)\right).\end{aligned}$$

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:

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

The next important question is how to we construct valid confidence intervals for the causal effect. It is a somewhat deep question because natural strategies (compute the variance of the estimator) will fail. We won't go into details here but instead we will discuss approaches that work.

28.7 Fisher's Exact p-values

Fisher (the father of MLE, Fisher's information and really much of modern statistics) 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.

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.

Of course, the sharp null is a very strong null hypothesis, and we often have much weaker null hypotheses. Like perhaps our null hypothesis is just that $\tau = 0$. We will discuss these in a future lecture.