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

Lecture 29: November 14

Lecturer: Siva Balakrishnan

29.1 Review and Outline

In the last class we discussed causal inference:

1. The potential outcomes framework
2. Causal estimands
3. The assignment mechanism, and randomized trials
4. An unbiased estimator for the average treatment effect
5. Fisher's exact p-values under the sharp null

Today we will continue our discussion of causal inference. In particular, we will focus on Neyman's version of things, particularly, on a new null hypothesis. Then we will turn our attention to observational studies.

Rather than rely on the sharp null, Neyman's idea for inference was: (1) derive the variance of the usual estimator of the average treatment effect, (2) estimate this variance, (3) appeal to the central limit theorem to construct confidence intervals.

29.2 A new causal estimand

In the last class, we assumed throughout that there was a fixed population of n individuals. Today, we will focus on the case where there is a super-population from which individuals are sampled i.i.d.

In this case, our causal estimand is the population average treatment effect:

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

where the expectation is over the distribution from which individuals are sampled.

29.3 Neyman's null hypothesis

Neyman was interested in testing a different hypothesis: whether the average treatment effect was zero or not, i.e.:

$$\begin{aligned} H_0 : &\quad \tau = 0, \\ H_1 : &\quad \tau \neq 0. \end{aligned}$$

We can use the same estimator:

$$\widehat{\tau} = \frac{1}{n_t} \sum_{i=1}^{n} \mathbb{I}(W_i = 1) Y_i(1) - \frac{1}{n_c} \sum_{i=1}^{n} \mathbb{I}(W_i = 0) Y_i(0).$$

Since our estimator $\widehat{\tau}$ is an average of i.i.d random variables. We know by the central limit theorem that:

$$\frac{\widehat{\tau} - \tau}{\sqrt{\mathrm{Var}(\widehat{\tau})}} \to N(0, 1),$$

as $n \to \infty$.

This means that we can test Neyman's null hypothesis by a Wald test, rejecting the null hypothesis if:

$$0 \notin [\widehat{\tau} - \sqrt{\mathrm{Var}(\widehat{\tau})} z_{\alpha/2}, \widehat{\tau} + \sqrt{\mathrm{Var}(\widehat{\tau})} z_{\alpha/2}].$$

The only remaining problem is then to compute/estimate the variance.

29.4 The variance of the average treatment effect

It turns out to be a fairly involved calculation, which we will not go through but is in the Imbens and Rubin book.

The variance is:

$$\mathrm{Var}(\widehat{\tau}) = \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t} - \frac{S_{tc}^2}{n},$$

where

$$\begin{aligned} S_c^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(0) - \mathbb{E}[Y(0)])^2, \\ S_t^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(1) - \mathbb{E}[Y(1)])^2, \\ S_{tc}^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0) - (\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]))^2. \end{aligned}$$

It turns out that we cannot estimate the third term because it involves terms like $\mathbb{E}(Y_i(0)Y_i(1))$ which we cannot estimate (even in a randomized trial). We can however ignore the third term and obtain an upper bound on the variance, which is

$$\mathrm{Var}(\widehat{\tau}) \leq \frac{S_c^2}{n_c} + \frac{S_t^2}{n_t}.$$

We can estimate this upper bound in a straightforward way. Particularly, focusing on the term, a natural unbiased estimate is:

$$\widehat{\sigma}_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n} \mathbb{I}(W_i = 0) \left(Y_i(0) - \frac{1}{n_c} \sum_{j=1}^{n} \mathbb{I}(W_i = 0) Y_i(0) \right)^2.$$

We can similarly estimate the other term and use this for inference.

29.5 Causal inference from observational data

A more difficult question that is the focus of most of the work on causal inference, is under what conditions can we use observational data (i.e. not data from a randomized trial) in order to estimate causal effects.

This is important because in most cases we have causal questions for which it is unethical to run a trial. For instance to answer the question: "does smoking cause cancer?" – the trial involved would need to randomly force people to smoke or not, which is obviously unethical.

On the other hand we have large amounts of observational data, i.e. data where there are many smokers and non-smokers, and a lot of health or "outcome" information.

The first problem is something called selection bias. Sticking with the population average treatment effect, a natural idea is to estimate this by taking the difference in outcomes over the smokers ($W_i = 1$) and non-smokers ($W_i = 0$):

$$\begin{aligned}
\widehat{\tau} &= \widehat{\mathbb{E}}[Y|W=1] - \widehat{\mathbb{E}}[Y|W=0] \\
&= \widehat{\mathbb{E}}[Y(1) - Y(0)] + \underbrace{[\widehat{\mathbb{E}}[Y(1)|W=1] - \widehat{\mathbb{E}}[Y(1)]] - [\widehat{\mathbb{E}}[Y(0)|W=0] - \widehat{\mathbb{E}}[Y(0)]]}_{\text{Selection Bias}}.
\end{aligned}$$

Roughly, selection bias is capturing the difference in potential outcomes amongst people who were observed to have been treated (or control) from the population potential outcomes, i.e. there is bias for instance if people who were going to get lung cancer anyway all decided to smoke.

How do we correct for selection bias? This is in general difficult/impossible.

Lets consider a simple scenario: suppose we have three variables: the treatment indicator W, the health outcome Y and income status X. In this example X is something we will call a confounder.

The reason we have selection bias, is that it is plausible that people with low income status, cannot afford good healthcare and are more likely to smoke. So we need to dis-entangle the "effect" of smoking on health outcomes from the effect of income status on health outcomes.

One idea is simply, to measure the income status, and then only compare people with the same (or similar) income status, i.e we can imagine comparing health outcomes amongst smokers and non-smokers for each income strata separately.

The general idea is to correct for so called confounders, i.e. find and measure a set of variables X which will make the potential outcomes independent of treatment, i.e.

$$Y_i(0), Y_i(1) \perp\!\!\!\perp W_i | X_i.$$

Now, lets suppose we can measure all the confounders, then we can observe that:

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

using the independence property above. This in turn means that:

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

and each of these terms can be estimated (with no bias). The usual way of estimating these is via regression, i.e. we estimate $\mathbb{E}[Y(1)|X, W = 1]$, via a regression of the covariates on the treated individuals and similarly for the controls. We then average these regression functions, and take the difference to estimate the causal effect τ.