Lecture 2 – Randomized Controlled Trials

Why Run Experiments?

Experiments can overcome the issue of selection bias. People don't get to decide if they are treated or not. This means that we identify the causal effect of treatment.

How does this work? It relies on the Law of Large Numbers. The Law of Large Numbers states that if we sample from a population, our sample average will converge to the population average as our sample grows.

If we randomize people in our sample into two groups, we expect that as we grow the samples the means will converge to the population means. This includes things that are easy to observe like gender or height but also things that are not easy to observe like how hard a student works or ability. With observable characteristics, we could just control for these using regression. However, the real magic of RCTs is that the LLN insures that *unobservable* characteristics will be balanced as well

We can test for this for easy to observe things like gender and height. We simply compare the mean in our two groups to each other. We can do this with a simple t-test or a regression. We expect that our samples will have the same average characteristics.

As a simple case, let's assume that $Y_{i1} = Y_{i0} + \kappa$. This is a constant treatment effect setting–everyone who is treated changes their outcome by κ.

For now assume people can decide to be treated or not.

$$E[Y_{i1} - Y_{i0}] = \kappa + E[Y_{i0}|D = 1] - E[Y_{i0}|D = 0] \tag{1}$$

Hence the expected difference among treated and control represents both the causal effect κ and the differences in the people who decide to be treated or the selection effect.

Now let's consider equation 1 if we have randomization. What do we know will be true? $E[Y_{i0}|D = 1] = E[Y_{i0}|D = 0]$ Why? Law of Large Numbers. Hence, comparing the group means in a randomized experiment reveals the true causal effect of treatment, κ This simple example extends to more complicated settings where there is not a constant treatment effect.

One reason to run experiments is that the can help us understand the true causal effect of treatment. That may excite economists but it may seem less appealing to other people. Randomization can be good for other reasons:

Fairness It may be that access to some program is limited due to funding (charter schools, Oregon medicaid expansion). In this case randomization has two benefits–easy evaluation and fairness.

Types of Validity

Internal validity for an experiment is about how well an experiment establishes the causal relationship. This means that randomization protocols were followed, balance on observables (and unobservables) was achieved, etc. A well run experiment should always be internally valid. Questions about internal validity include: Did subjects understand what they were supposed to when they were in the lab? Were the rules about randomization followed? Was data collected in a reliable way? etc.

External Validity is how generalizable the result is. Do the results of this study apply to other settings? Let's say that UT Austin runs an experiment and gives out more financial aid to some students and finds that this does not affect graduation. Does this tell us what is likely to happen at a community college? UT San Antonio? A for-profit college? This is the issue of external validity

Returning to the question of admission to a charter school: Charter school admissions are often internally valid because randomization procedures are followed. However, if only very high ability students sign up to attend a charter school then the experiment is unlikely to be externally valid.

Assumptions

Stable Unit Treatment Value (SUTVA)–Treatment only affects a single unit.

This is violated if there are spillovers from treated people to untreated people. For instance, let's say you randomly treat half of a classroom with tutors and half does not receive tutors. There may be spillovers where the tutored kids help the kids without tutors to learn. If this is true, would comparing the treated student's performance to the untreated students performance uncover the causal effect of tutors? Would it likely be an underestimate or an overestimate of the true treatment effect?

This assumption should be carefully considered when designing an experiment. Should treatment be at the person level or the classroom level?

Estimation

How do we estimate treatment effects in RCTs?

Essentially we just want to compare means. We could do this with a t test.

We can also use regression to compare means. We can use the following equation:

$$Y = \beta_0 + \beta_1 D + \varepsilon \tag{2}$$

β_0 tells us mean for the control group. β_1 tells us the treatment effect. Regression is nice because we could add controls for predetermined covariates. We only want to

control for things that occur *prior* to treatment, that cannot be affected by treatment.

Do we need to add controls at all given what we talked about with the LLN? No, observed and unobserved should be balanced. However, people do often control for covariates because it reduces the standard errors. This is because there will be small (likely statistically insignificant) differences in predetermined covariates.

One of the important things to check for in an RCT is that predetermined covariates are "balanced." This means that treatment and control groups look similar. You can do this by replacing Y in the equation above with X

Interpretation

Proper interpretation of RCTs is key.

Researchers are constrained by who signs up for an experiment. For instance, using randomization of admission to charter schools can tell us about the effects of charter schools *among students who attempted to register*. We cannot know what it would do for students who did not sign up to go to a charter school without addition assumptions. This is a statement about external validity.

Defining treatment–Treatment is often a bundle of things. For example, you get additional financial aid for college AND counseling. You cannot separately recover the effect of aid and counseling. If instead, people were randomized into receiving aid and separately into receiving counseling, you could identify the separate effects. Some things in the bundle are more subtle–for instance some people may get access to a government program but also must fill out more paper work.

We cannot know the causal effect of things that change as a result of treatment that are not randomly assigned. For instance, supposed you run an experiment where you find that a job training program increases earnings and classes taken a community college. You cannot say that the increase in earnings is because of the classes taken; however, you can say that treatment increases both classes taken and earnings.

I often summarize this by saying, "You can only know the effect of the thing that was randomized."

The time frame of the analysis is important. Moving to Opportunity (MTO) is a program that gave low income families vouchers (and other things) to help them move into higher income neighborhoods. Early evaluations of this work generally found no effects. However, considering the longer term outcomes and focusing on the youngest kids revealed that there was an effect. (Chetty Hendren Katz 2016)

Social Phenomenon in RCTS:

Hawthorne effect People may change their behavior because they know an experiment is happening.

Experimenter Demand Effect People may try to behave in a way that will make the researcher happy

Placebo Effect In drug trials, sugar pills can have a positive effect on health. You may have similar things happen in social science settings. One way to combat this is to make the control condition similar to treatment.

Multiple Testing If you run an RCT you need to be very careful about multiple testing or "p-hacking." If you run an RCT and there was no effect, you can still get spurious effects if you perform many tests. For instance, if you split the sample into 20 different groups, you would expect one of the groups to have an effect significant different from zero by chance. This is because we typically construct 95% confidence intervals. Another way you can get in trouble is by looking at many outcomes which also increases the number of tests you run.

To combat this there are a few things you can do:

Use theory to tell you where you expect to find effects. Don't run a bunch of specifications and then decide to come up with your theory.

Pre specify your analysis plan. This is an increasingly common practice. The American Economic Association maintains a website where researchers pre speficy their analysis plan. Researchers discuss what outcomes they are interested in, what populations, and how they will deal with multiple testing https://www.socialscienceregistry.org/

Use a multiple testing correction. There are several–the most extreme and earliest is Bon Ferroni

Use common sense–don't torture the data to make your preferred story appear.

Types of RCTs

Lab Experiment Subjects are recruited and brought into a lab. In the lab some sort of experiment is performed with outcomes measured in the lab. Pros: comparatively cheap and convenient, able to control most of the setting. Cons: external validity

Lab in Field Subjects are recruited in a particular setting that may be of interest (people from low-income country). Similar to lab experiments

Field Experiments Experiments done in a real world setting. Pros: better external validity–the stakes are real Cons: much harder to implement, costlier

Sometimes researchers randomize using more complex rules. For instance, there may be several treatment arms. There is also stratified random design which ran-

domizes withing a group. Say you want to make sure you have the same number of men in treatment and control. You can randomize among men and among women. If you do not do this, you may have situations where not very many of a certain type of observation are in treatment or control. This is important for RCTs with small numbers of observations.

Power and Inference

Another consideration is statistical power. Statistical power is

$$power = Pr(\text{reject} H_0 | H_1 \text{is true}) \tag{3}$$

Tests with high power mean that they are likely to to reject H_0 when H_1 is true. Power depends on the underlying distribution of the variable, the size of the effect, and the type of test.

Before engaging in an expensive RCT researchers often (should) do power calculations. These require assumptions about the size of the potential effect and the variance of the outcome. If you have these, you can compute how large a sample is needed. Typically researchers use a standard of .8 for power; that is, if there is an effect they will reject the null of no effect 80% of the time.

See http://powerandsamplesize.com/ for help in power calculations.

Suppose you run an experiment where 30 hospitals are allocated to treatment and control (15 each) and you're interested in an experiment designed to reduce wait times in the emergency room. You get data from each of the hospitals on how long each patient waited to be seen. Each hospital saw 1,000 patients so you have data from 30,000 patients. It might be tempting to compare the mean differences in wait time using individual data. However, this will give you the wrong standard errors. Because treatment assignment is at the hospital level (not the patient level) you need to adjust your standard errors. Functionally you have 15 treated units and 15 control units–not 15,000 and 15,000. We will return to this later in the class.

RCTs are very simple–we essentially just compare means. We can do this using a t-test. Often, we use regression to accomplish the same thing. We get exactly the same answer if we do not use covariates. However, we often control for covariates to test for mean differences. This accounts for small differences that may arise by chance across our treatment and control and generally increases our statistical precision.

Ethical Issues

Some experiments are unethical. Historically, researchers have run unethical experiments including Nazi human experiments, Tuskegee syphilis experiment, and others.

As a result, Institutional Research Boards (IRBs) now approve research that involves human subjects. If you are running an experiment with people (or some other types of research) a university IRB should approve your experiment prior to your research.

Economics employs a "no deception" policy. This means that researchers cannot lie to subjects and makes it so researchers are credible in the long run.