PPOL 503-03, PPOL 503-04, Fall 2016
Course Notes #13: Experiments and Quasi-Experiments

I. Framework of Randomized Experiments

1. Existing evaluations focus almost exclusively on estimating mean impacts. The difficulty with estimating mean impacts of a program arises in constructing the desired counterfactual. The counterfactual is what would have happened to the treatment group in the absence of the program.

2. Consider a voluntary program. To estimate the mean impact of participation on participants, we need an estimate of the mean outcome that would have been obtained had the participants not participated. Recall that a key concept in evaluating the *impact* (i.e., cause) of a program (i.e., treatment) is the *counterfactual*. Using the notation found in Heckman and Smith (1995), the goal is to estimate $E(Y_1|d=1) - E(Y_0|d=1)$, which is the mean causal impact of a treatment program on the participants.

3. Need to estimate $E(Y_0 \mid d = 1)$. This represents the counterfactual. It is the mean outcome that would have been obtained had the participants not participated in the treatment. This is impossible to know.

4. Instead, we can estimate $E(Y_0|d=0)$ as a proxy for the counterfactual. This is the mean outcome of those who do not participate in the treatment. Is this a good proxy for the counterfactual? The answer is NO.

5. We cannot use the mean outcome among nonparticipants $E(Y_0 \mid d = 0)$ as a proxy for what would have happened to participants had they not participated.

$E(Y_1 \mid d = 1) - E(Y_0 \mid d = 0) =$

$\{E(Y_1 \mid d = 1) - E(Y_0 \mid d = 1)\} +$
(program effect)

$\{E(Y_0 \mid d = 1) - E(Y_0 \mid d = 0)\}$
(selection effect – due to the fact that nonparticipants differ from participants in the nonparticipant state).

The selection bias term does not equal zero.

6. Randomized social experiments solve the problem of selection bias by generating an experimental control group composed of persons who would have participated but who were denied access to the program or treatment.

Let r = 1 denoting randomization into the treatment group, and let r = 0 denote randomization into the control group, which is denied access to the treatment. Let d* = 1 for persons who would participate in a program in the presence of randomization and d* = 0 for everyone else.

$E(Y_1 - Y_0 \mid d = 1) =$

$E(Y_1 \mid r = 1 \text{ and } d^* = 1) - E(Y_0 \mid r = 0 \text{ and } d^* = 1)$

Mean of Experimental Mean of Control

7. Randomization acts as an instrumental variable by creating variation in the receipt of treatment among participants.

8. Randomization does not remove selection bias but rather balances the bias between participant and nonparticipant samples. Nothing guarantees that the expected value of the disturbance equals zero.

Rather, randomization balances the bias in the two samples so it cancels out when estimating the mean impact.

9. Note that Heckman and Smith (1995) argue that non-experimental (observational) studies can construct valid counterfactuals econometrically. We will study some of these methods later in the course. However, it is commonly (though not universally) accepted that these methods are more difficult and rely on stronger assumptions (though the results might have greater external validity). Freedman (1991) makes the case that econometric techniques are not adequate substitutes for a good research design (using "shoe leather").

10. One can say the same thing using Burtless's (1995) notation. Assume that the *true* model is

$$Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon.$$

Selection bias results if there are some unobservable determinants of Y (e.g., if we cannot observe X then it is part of the disturbance term) that are correlated with selection into the treatment group (T). Give some examples of why this would happen?

11. How does randomization address the selection problem described in the previous equation?

With randomization the *distribution* of the two groups on all baseline characteristics (observable and unobservable) will be identical up to a known degree of sampling error. One exception: participation in the program. Note the importance of sample sizes. Note also Heckman and Smith's additional "assumptions … that randomization does not alter the pool of participants or their behavior and that close substitutes for the experimental treatment are not readily available."

12. Basic advantage of experiments: Because members of the groups (treatment and control) do not differ systematically at the outset of the experiment, any difference that subsequently arises between them can be attributed to the treatment rather than to other factors.

13. The essence of a randomized experiment is thus to compare the mean outcome of the treatment group with mean outcome of the control group.

14. Two key concepts:

- Effect of Intention to Treat (ITT): Refers to the impact of the program on those who were assigned to the treatment group. That is, ITT measures the impact of the *offer* to participate in the program. Note that not all members of the treatment group necessarily received the treatment (i.e., not everyone randomly assigned to participate in the program actually participates).

- Effect of the Treatment on the Treated (TT): Refers to the mean impact of the program on those members of the treatment group who actually participated in the program.

15. Can you estimate the impact of the treatment on the treated (TT) by comparing the mean outcomes of the participants with those of the control group?

- Random assignment creates comparability between the *entire* treatment group and the *entire* control group; this strength of experimental methods does not *necessarily* apply to subgroups of the treatment and control groups.

- In general, it is not possible to derive experimental impact estimates on *endogenously* defined subgroups (i.e., groups defined on the basis of events or actions that occur after random assignment).

• Note also that the causal effect is likely to be different for different people in the treatment group. We are estimating the *average causal effect*.

II. NUTS AND BOLTS OF CONDUCTING A RANDOMIZED EXPERIMENT

1. Design the study carefully. Do you have large enough sample sizes to detect policy-relevant impacts?What are your outcome measures?

2. Collect baseline data and randomly assign people to the treatment and control groups
3. Verify that assignment looks random (comparison of observable means)

4. Monitor process so that integrity of experiment is not compromised:

• No-shows: members of the treatment group who do not receive the treatment.

• Crossovers: members of the control group who receive the treatment.

5. Collect follow-up data for both the treatment and control groups. That is, collect data on *outcomes* some time after the program was implemented.

6. Estimate program *impacts* by comparing mean outcomes of treatment group vs. mean outcomes of control group.

- Not using statistical controls
  - Impact = $E(Y_T)$-$E(Y_C)$, where $E(Y_T)$ is the mean outcome of the treatment group and $E(Y_C)$ is the mean outcome of the control group.
- Using statistical controls (i.e., "regression-adjusted" or "differences estimator with additional regressors")

Run regression of
$Y = \beta_0 + \beta_1 T + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$, where Y is the outcome variable, T is a dummy variable indicating whether the person is in the treatment group, and the X's are observable characteristics at baseline.
β_1 represents the impact of the program on the outcome Y.

Why is β_1-*hat* is unbiased? This is due to randomization.

Discuss why to include regression-adjusted estimates. 1) Efficiency; 2) a large discrepancy with non-regression adjusted results suggests that random assignment failed.

III. PROBLEMS WITH EXPERIMENTS

1. A statistical study is internally valid if the statistical inferences about causal effects are valid for the population being studied. A study is externally valid if its inferences and conclusions can be generalized from the population and setting being studied to other populations and settings.

2. Threats to internal validity include: failure to randomize, failure to follow the treatment protocol, attrition, experimental effects, cross-over bias and small sample sizes.

a) Failure to randomize. If the treatment is not assigned randomly, but instead is based on characteristics or preferences of the subject, then experimental outcomes will reflect both the effect of the treatment and the effect of the nonrandom assignment. Example: suppose that participants in a job training program are assigned to the treatment group depending on whether their last name falls in the first or second half of the alphabet. Because of ethnic differences in last names, ethnicity could differ

systematically between treatment and control groups. Thus, nonrandom assignment can lead to correlation between the treatment and the error term because receiving the treatment is determined in part by individual characteristics that enter the error term.

b) Failure to follow the treatment protocol. Even if the treatment assigned is random, the treatment actually received might not be random. Instead, the treatment the subject actually receives is partly determined by random assignment and partly by the individual's characteristics (the subject's desire to receive the treatment). You can require a student to take a course, but it is more difficult to ensure class attendance.

c) Attrition refers to subjects dropping out of the study after being randomly assigned to either the treatment or control group. If the reason for attrition is related to the treatment itself, then attrition results in bias in the OLS estimator of the causal effect. Example: suppose the most able job trainees drop out of the program experiment because they obtain out-of –town jobs using the skills acquired during job training program. Only the least able members of the treatment group remain at the end of the experiment. If so, the treatment will be correlated with the disturbance for those who remain in the sample at the end of the experiment. Because attrition results in a nonrandomly selected sample, attrition that is related to the treatment results in selection bias.

d) Experimental effects. Sometimes the mere fact that individuals are in an experiment can change behavior, a phenomenon known as the Hawthorne effect. In some experiments, a "double-blinded" protocol can mitigate the effect of being in the experiment. Although subjects and experimenters know that they are in an experiment, neither knows whether the subject is in the treatment or control group. (medical experiments with drug and placebo)

Double-blind experiments are infeasible in the real world. In a poorly designed experiment, the experimental effect could be substantial. For example, teachers in the experimental program might work extra hard to make the program a success if they run the risk of losing their job.

e) Substitution (or Crossover) Bias – Can members of the control group obtain close substitutes for the treatment elsewhere? If so, then the results will be biased downwards. You can adjust for this if you know what proportion of the control group "crossed over," and if you can assume that the program had the same effect on crossovers that it would have had if they had been assigned to the treatment.

f) Small samples. A small sample size does not bias estimators of the causal effect, but it does mean that the causal effect is estimated imprecisely.

3. Threats to external validity. Compromise the ability to generalize the results of the study to other populations.

a) Nonrepresentative Sample. The population studied and the population of interest must be sufficiently similar to justify generalizing the results. Selecting the sample nonrandomly from the population of interest can compromise the ability to generalize the results from the population studied (such as volunteers) to the population of interest.

b) Nonrepresentative program or policy. The policy or program of interest must sufficiently similar to the program being studied to permit generalizing the results. The program in a small-scale tightly monitored experiment could be quite different than the program actually implemented. Another difference could be duration:

suppose the experiment lasts only a few months while the actual program is available for longer periods.

c) General equilibrium effects. Turning a small, temporary experimental program into a widespread, permanent program might change the economic environment sufficiently that the results from the experiments cannot be generalized. For example, a widespread educational reform such as school vouchers or sharply reducing class size, could increase the demand for teachers and change the type of person who is attracted to teaching. The net effect of the reform would reflect these induced changes in school personnel. A small experiment might hold these environmental effects constant, but it may be impossible to hold such general equilibrium effects constant when the program is broadly implemented.

d) Treatment versus Eligibility Effects.

Voluntary participation provides a biased estimator of the program effect. Examples: evaluating the effects of participation in a reading program for third grade students in DC on reading test scores. Participation is voluntary and this must be accounted for in evaluating the effects of the reading program on test scores.