Problem Set 1

1 Purpose

The purpose of this problem set is to assess your understanding of one key method of quantitative public opinion research: experimental design and analysis.

2 Your Task

1. In your own words, explain the "potential outcomes framework" of causal inference and explain how experiments provide a way to identify causal effects.

   Solution:

   The potential outcomes framework (or Neyman–Rubin–Holland causal model) asserts that for each unit, there exist (in theory) one or more counterfactual "potential" outcomes, only one of which is *realized*. Causality is defined as the difference in an individual unit's potential outcomes. Because only one potential outcome can be realized, we cannot observe causality at the unit-level.

   There are then multiple ways to think about how experiments identify causal effects.

   - Randomized experiments do not allow us to see individual-level causal effects unless we assume unit homogeneity. In all other cases, randomization allows us to assess average causal effects of a treatment on an outcome.
   - Randomized experiments randomly sample potential outcomes in an unbiased manner, allowing for estimation of an average treatment effect.
   - Randomized experiments randomly expose one of multiple potential outcomes for each individual in the sample.
   - Randomized experiments eliminate confounding.
   - Randomized experiments eliminate selection bias into treatment assignment because the we examine variation in an outcome due to variation in treatment induced by the randomization process.

2. A researcher wants to understand how the provision of cash incentives to poor families with children affects the educational attainment of the children and considers two alternative designs.

   (a) The first design involves examining the educational attainment of children whose parents are or are not eligible for public cash assistance by generating a random sample of the population, selecting individuals with incomes just above and just below the cutoff for assistance, and tracking the educational progress of their children.

(b) The second design involves recruiting a non-representative sample of families that are not currently eligible to receive benefits (but are close to eligibility). One half of families are randomly assigned to receive cash assistance and the other half is randomly assigned to receive nothing. Educational attainment is tracked for both groups.

Discuss the trade-offs involved in these designs, including what would be required to obtain an estimate of the causal effect of cash assistance on educational attainment.

Solution:

There are numerous trade-offs here and there is no correct answer. Some trade-offs you could consider include:

- The first design uses a population-based (i.e., representative) sample but includes no experimental component. The representativeness ensures a degree of external validity because the group of participants are selected from (a subset of) the population as a whole. The second design is an experiment involving participants recruited from a convenience (i.e., non-representative) sample, thus potentially limiting external validity because those who are locally convenient may differ from the population of people who are just above the threshold nationwide.
- Strictly speaking, the two designs estimate effects on slightly different populations: the first for the population just above and just below the cutoff; the second only for the population just above the cutoff.
- The first design is a "regression discontinuity design" where it is assumed that those just above and just below the cutoff are sufficiently similar to provide counterfactuals for one another. But, given that the cutoff already exists, one needs to be cognizant of "Campbell's Law" (i.e., the exploitation of social indicators, which induces selection pressures and/or compensatory behaviour). We may also then need to introduce controls for any factors that differ between those just above and just below the cutoff.
- The second design is a straightforward experiment, so estimation of effects is simple provided that there is compliance with the treatment (i.e., no one receives a treatment other than what they are assigned) and no attrition (i.e., no one leaves the study). There may be ethical questions related to providing assistance to a random subset of people that would be worth considering.
- Both designs suffer from ambiguity about how outcomes (i.e., educational attainment) will actually be measured. This could be quite complex given that education is a long-run outcome that may not be known for many years. This invites further complexities about deciding when to stop the study, the cost of monitoring outcomes, etc.
- Both studies will have similar power — a sample of roughly equal size in either design will generate approximately similar variances.

3. Consider an experiment on 500 individuals in which one group is randomly assigned to read a treatment message from Boris Johnson supporting a "no deal" position in the ongoing

Brexit negotiations and another group is assigned to a control condition that receives no information. Measures of opinions for about support for "no deal" are recorded for both groups on a 0 to 1 scale, with higher scores indicating greater favorability toward "no deal."

(a) Assuming the treatment group mean score was 0.68 and the control group mean score was .51, what is the average treatment effect? Is this substantively large or small?

> **Solution:** The sample average treatment effect is simply the mean difference: $0.68 - 0.51 = 0.17$.
>
> Deciding whether this is large or small requires some kind of benchmark for comparison. We could say this is 17% of the scale (from 0 to 1). Or, we could compare it to the standard deviation of the outcome to express a "standardized mean-difference" but that information is not provided here. Assuming the standard deviation were 0.4, then the effect size would be $\frac{0.17}{0.40} = 0.425$, and so forth. The decision about whether an effect is substantively large or small is subjective depending on what the researcher (and consumers of the research) would consider to be large or small.

(b) Assuming the t-statistic for the mean-difference is 1.76, should we consider this effect to be statistically large and distinguishable from zero?

> **Solution:** Answering this question technically requires consulting a t-distribution. You can find one on Wikipedia. Given the very large sample size, use the ∞ row of the table. a t-statistic of this size is considered statistically distinguishable from zero in the case of a one-tailed test ($p < 0.05$) but not in a two-tailed test ($p < 0.10$). A very simple rule-of-thumb is that a t-statistic larger than 2 indicates that the difference can be considered statistically significant in a large sample, for a two-tailed test. (If we had a strong directional prediction that the treatment outcome would be higher than the control outcome, we could consider the one-tailed test appropriate but probably not otherwise.)

4. What is a randomisation distribution? What can we learn from the randomisation distribution for the sample average treatment effect?

> **Solution:** The randomisation distribution is a particular kind of sampling distribution for the average treatment effect, which conveys the variation in estimates of an ATE (that is actually 0) due to chance variation. Rather than assuming a "parametric distribution" (such as the t distribution), the randomisation distribution uses the data we observe from our experiment to create a set of possible values of the ATE that we could see in these data if the true ATE were 0 (i.e., there is no difference between the two treatments) and we simply shuffled which units were assigned to which treatment, calculating the apparent ATE in each re-randomisation.
>
> The randomisation distribution allows us to see whether our observed ATE is unusually large or small (i.e., to calculate a p-value) as we all as calculate the variance or standard error of the ATE without assuming a particular parametric distribution. In large samples, this will be identical to values generated from a t-distribution.

5. The statistical power of a two-sample t-test (which is, in essence, the power of a posttest-only, two-group experimental design) is influenced by four main things: the size of each experimental group, the difference-in-means (i.e., difference in mean values of the outcome in the two groups), the variance of the outcome measure, and α (the significance level or "Type 1" error probability).

   (a) If α (the Type 1 error probability) is 0.05, how often should we expect to find a "statistically significant" effect size when one is not present?

   Solution:
   This is simply 0.05 or 5% of the time (i.e., the "false positive rate").

   (b) If you increase the size of your treatment groups in an experiment while the expected effect size remains unchanged, what happens to the power of your experiment? Are you more or less likely to obtain a "false zero" result? What about "false positives"?

   Solution:
   The power of the test increases. Recall the definition of power:

$$Power = \phi\left(\frac{|\mu_1 - \mu_0|\sqrt{N}}{2\sigma} - \phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right) \tag{1}$$

   Without worrying about all the details of the equation, note that sample size is in the numerator of the first term, so more observations means more power. This directly translates into a lower likelihood of false zeroes $(1 - \beta)$ where power is denoted β.
   This is unrelated to the false positive rate, which is a function of the selected significance level, α, only, so we would still expect to encounter false positives 5% of the time.

   (c) Imagine we are expecting to find a small effect but we can only collect a small number of observations in our experiment, so the minimum detectable effect size in our study is larger than the effect size we would expect to observe given our theory. If our experiment reveals an effect that is statistically distinguishable from zero, what are the two possible interpretations of this result?

   Solution:

   i. The effect is actually larger than we expected.

   ii. The effect in our experiment is a massive overestimate (i.e., a false positive).

   We cannot distinguish which alternative is correct.

6. Sometimes experiments are "broken" due to challenges of implementing an intervention and measuring outcomes. In what ways can experiments fail? And what implication(s) does each of those points of failure have on the analysis of the experimental data and the interpretation thereof?

Solution:

- Noncompliance: sometimes units do not receive the treatment they were assigned. This could be due to administrative failures in implementing the intervention or because of self-selection out of treatment. This limits our ability to understand the effect of treatment in potentially complex ways. If noncompliance is asymmetric (i.e., occurring only in one group), our effect size is diminished. If noncompliance is symmetric (i.e., occurring in two or more groups), we do not know the direction of bias.
- Attrition: sometimes units leave the study before it has finished. This limits our ability to understand the effect of treatment both by reducing effective sample size (we have less power than anticipated) and by potentially biasing treatment effects if attrition is nonrandom (i.e., correlated with pretreatment covariates or treatment itself).
- Missing outcome data: sometimes we simply have missing data on outcomes (e.g., due to failure to measure or record outcome data). While potentially distinct from attrition, the implications are the same (reduced power and potentially biased treatment effect estimates).
- Contamination/Spillover: sometimes units are not independent but instead are networked in some way (e.g., students within classes or families within households). If we attempt to closely linked units to distinct treatments, we have to be aware that one treatment might bleed into or contaminate another unit's experiment. This is easily avoided by, for example, not assigning units in a household to different conditions. If it occurs, we violate a core assumption called SUTVA (stable unit treatment value assumption) and largely cannot know what effect treatment had on units without strong assumptions about how units are interconnected.
- Treatment heterogeneiety: sometimes units in the same experimental condition in practice receive different treatments (e.g., variations on a conversation, variations on an education intervention, etc.). This may or may not be problematic. If our goal is to understand in broad terms what the effect of an intervention is, this is not a problem. If we care about mechanisms or the specific details of what part of the intervention affected outcomes, then units are actually not providing consistent evidence about a single intervention and we therefore learn nothing.

7. Are experiments more or less useful than other methods for generating evidence-based policy? Why? What caveats — if any — should be placed on the use of experimental evidence for policy and decision-making?

Solution: This is completely subjective. We have discussed lots of advantages and disadvantages of experimentation for providing causal evidence. Experiments provide clear causal inference but can only be used in certain circumstances. They may trade-off internal validity and external validity and they may pose ethical issues or constraints. Whether experiments should inform public policy and do so more than other types

of evidence therefore requires judgements about the competence of policy-makers to interpret different kinds of evidence.

3 Submission Instructions

Please submit your answers as a PDF document via Moodle. It should be no more than 4 pages, single-spaced, in Times New Roman font size 12, on A4 paper with standard 2.54cm margins. The code for R or Stata to reproduce results should be included as an appendix, written entirely in fixed width format font (e.g., Courier New). A solution set will be provided on the course website and the activity will be discussed in class.

4 Feedback

Group feedback will be provided during class. If you would like more specific individual feedback on your work, please ask the instructor during class or office hours.