Problem Set 3: Experimentation

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 Overview

You are asked to watch a short slidecast covering relevant material and then complete the following tasks. Questions regarding this material should be raised on the Moodle discussion board (`https://moodle.lse.ac.uk/mod/forum/view.php?id=521541`) or in instructor office hours.

3 Your Task

1. In your own words, explain the "fundamental problem of causal inference" and how experiments provide a solution to that problem.

   Solution:

   There are multiple ways to think about this.

   - 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 selection bias into treatment assignment.
   - Randomized experiments eliminate confounding.

2. A researcher wants to understand how a televised party leaders debate affected citizens' vote intentions and considers two alternative research designs. The first design involves interviewing a representative sample of citizens, asking whether they watched the debate, and comparing vote intentions among viewers and non-viewers. The second design involves recruiting a non-representative sample of citizens into a laboratory session where one half of participants are randomly assigned to watch the debate and the other half is randomly assigned to watch a non-political program. Vote intentions are measured at the end of the laboratory session. Discuss the trade-offs involved in these designs, including what would be required to obtain an estimate of the causal effect of the debate on vote intentions in each design.

Solution:

There are numerous trade-offs here and there is no correct answer. The first design is a population-based survey with no experimental component. The second design is a laboratory experiment involving a non-representative sample.

The former design has claims to "external" validity because sample characteristics are unbiased estimates of population parameters. In attempting to infer causation, however, the design does not necessarily address selection bias: it does not induce debate viewing, so we cannot know why people watched the debate and whether any of those factors also explain vote intentions (e.g., some third variable explains both forms of political engagement). The measurement of vote intention occurs some time after the debate, possibly suggesting durability of any influence.

The latter design has claims to "internal" validity because randomized assignment to debate viewing enables an estimation of an average causal effect of viewing the debate but that effect does not necessarily reflect the average effect in any population. The measurement of vote intentions occurs immediately. In both designs, this measurement of vote intention may not say anything about actual voting behavior.

Both designs likely present other trade-offs: feasibility, cost, etc.

3. How do we know if randomization "worked"? In other words, how do we know that experimental groups are identical to one another except for the difference in the experimentally manipulated variable?

Solution: There are two basic answers:

(a) We don't. If we used a physical randomization process (i.e., something we "know" is truly random), then we do not check anything. Randomization creates balance in expectation and we are using a randomized process.

(b) Balance tests. We select a set of covariates that we think are important to have balance on and we perform something akin to a t-test for each variable using the treatment assignment variable as a right-hand side variable. If non of these tests indicate a significant difference in covariates, then the covariates are balanced and we say treatment randomization "worked."

The first appeals to "design-based" inference while the latter assumes that we have to demonstrate successful randomization rather than assume it by design.

4. Consider an experiment on 500 individuals in which one group is randomly assigned to read a treatment message from David Cameron support the "Remain" position in the upcoming European Union referendum and another group is assigned to a control condition that receives no information. Measures of opinions for the European Union are recorded for both groups on a 0 to 1 scale, with higher scores indicating greater favorability toward British membership in the European Union.

   (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.

(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 ($p < 0.10$). 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.

(c) Is the effect of this treatment larger or smaller than one from an earlier study with an effect size of Cohen's d = 0.40?

Solution:
Answering this question requires knowing the relationship between the mean-difference, Cohen's d statistic, and the t-statistic. The mean-difference is simply the difference between each treatment group mean. Cohen's d is simply the mean-difference over the pooled standard deviation of the outcome. The t-statistic is the same quantity over the standard error of the mean-difference.
To convert t to d, simply calculate: $\frac{2t}{\sqrt{df}}$, where df is degrees of freedom, which is simply $n_1 + n_2 - 2$. In this case, this is simply:

$$d = \frac{2t}{\sqrt{df}} = \frac{2 * 1.76}{\sqrt{500 + 500 - 2}} = 0.11 \tag{1}$$

In this case, the effect size in our experiment is much smaller than the effect size in the earlier study (0.40). Indeed, it is about 28% as large.

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 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 alpha (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.

(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 negative" result? What about "false positives"?

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

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

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 negatives $(1 - \beta)$ where power is denoted β.
Technically, this is unrelated to the false positive rate, which is a function of the selected significance level, α, only.

(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.
>
> We cannot distinguish which alternative is correct.

6. Consider an experiment in which the effect of a treatment is measured using a single survey question. Assuming a given effect size, and that we cannot change α or the size of the experimental groups, what practical action can we do to increase the power of our experimental design?

> **Solution:**
>
> The only thing we can change in this design is the variance of the outcome. One's immediate reaction might be that this is impossible, but that is not true. The easiest way to reduce the variance of an outcome is to create a more precise measure. That can be achieved by using multiple rather than a single item measure, creating a simple scale.
>
> If it is cheap to add additional questions (and it might not be, given the particular design), then adding questions in this way might be much more affordable than increasing sample size, given the dramatically declining marginal power of increasing sample size.

4 Submission Instructions

You should submit your problem set as a Word (.docx) document or PDF via Moodle.

5 Feedback

You will receive feedback within two weeks.