PLSC 30600: Problem Set 1

YOUR NAME

January 3, 2024

This problem set is due at **11:59 pm on Tuesday, January 16th**.

Please upload your solutions as a .pdf file saved as "Yourlastname_Yourfirstinitial_pset1.pdf"). In addition, an electronic copy of your .Rmd file (saved as "Yourlastname_Yourfirstinitial_pset1.Rmd") must be submitted to the course website at the same time. We should be able to run your code without error messages. In order to receive credit, homework submissions must be substantially started and all work must be shown. Late assignments will not be accepted.

Problem 1

In this problem we will examine what information the data might provide us regarding the magnitude or direction of a treatment effect if we are only willing to make a consistency or SUTVA assumption with respect to the potential outcomes and a positivity/overlap assumption on the probability of treatment.

Consider our standard causal inference setup with a binary treatment and a binary outcome. Y_i denotes the observed outcome for unit i, $Y_i \in \{0, 1\}$. D_i denotes the observed treatment for unit i. $Y_i(d)$ denotes the potential outcome we would observe if i were assigned treatment value d. By consistency, we have: $Y_i(d) = Y_i$ if $D_i = d$. By positivity, we have $0 < Pr(D_i = 1) < 1$. In other words, treatment is not deterministic and each unit could have received either treatment or control.

We'll focus first on bounding the average potential outcomes under treatment: $E[Y_i(1)]$

Part A

Write $E[Y_i(1)]$ in terms of the difference between the observed treated group mean $E[Y_i|D_i = 1]$ and a bias term. Hint: Use Law of Total Expectation.

Part B

Which components of the bias term can be observed from the data and which ones cannot? Can $E[Y_i(1)]$ be point-identified from the data alone? Explain why or why not.

Part C

What is the smallest possible value of the bias term? What is the largest possible value of the bias term? (remember that Y_i is a binary value that can either be 0 or 1).

Part D

Using your result from C, find the upper and lower bounds on $E[Y_i(1)]$.

Part E

Use the same approach from Parts A-D to find the upper and lower bounds on $E[Y_i(0)]$. Can you conclude anything about the sign of the average treatment effect $\tau = E[Y_i(1)] - E[Y_i(0)]$ from SUTVA and positivity alone?

Hint: Using the two sets of bounds, what's the largest possible treatment effect? What's the smallest possible treatment effect? Do these have the same sign?

Problem 2

Can conversations change minds on political issues? Broockman and Kalla (2016) attempted to answer this question by conducting a door-to-door canvassing experiment in Miami, Florida that followed in the wake of a Miami-Dade County transgender anti-discrmination ordnance that was passed in 2014. Partnering with a local LGBT organization, the study designed and evaluated a door-to-door canvassing intervention in which canvassers engaged with voters on transgender rights using a "perspective-taking" conversation strategy.

68,378 registered voters were recruited to take a baseline survey. Of these, 1825 completed the baseline survey and were selected to be assigned to either the treatment group receiving the transgender canvassing treatment or to a control group which received a "placebo" canvassing conversation about recycling. After the intervention, individuals who answered the door were recruited to complete a series of follow-up surveys – at 3-days, 3-weeks, 6-weeks and 3-months.

The complete citation for the study is below

> Broockman, David, and Joshua Kalla. "Durably reducing transphobia: A field experiment on door-to-door canvassing." Science 352, no. 6282 (2016): 220-224.

For this problem, you will need the `broockman_kalla_replication_data.dta` file. The code below will read in the file.

```
canvas <- haven::read_dta("broockman_kalla_replication_data.dta") %>% filter(!is.na(treat_ind))
```

Five surveys were conducted in total: - Survey 0 is the baseline survey - Survey 1 is conducted 3 days after the intervention - Survey 2 is conducted 3 weeks after the intervention - Survey 3 is conducted 6 weeks after the intervention - Survey 4 is conducted 3 months after the intervention

The variables you will need are:

- `treat_ind` - Indicator for which treatment was received by the respondent (1 = treatment (transgender rights), 0 = control (recycling))
- `contacted` - Indicator for whether the respondent was successfully contacted by the canvasser.
- `hh_id` - Unique identifier for each household.
- `respondent_t#` - Did the individual respond to survey #?
- `therm_trans_t#` - Feeling thermometer towards trans people (0-100) # denotes the response in a given survey wave (0, 1, 2, 3, 4). Positive values denote more favorable attitudes.
- `vf_age` - Age of respondent (from the voter file)
- `vf_party` - Party ID of the respondent (from the voter file)
- `vf_racename`- Race of the respondent (from the voter file)

Part A

Start by evaluating the covariate balance between households assigned to treatment and those assigned to control on the three sets of baseline covariates: age, party ID and race/ethnicity (all obtained from the voter

file). Do respondents assigned to receive the transgender canvassing intervention differ noticeably in these characteristics compared to those assigned to the placebo?

Part B

Treatment was randomly assigned at the level of the **household** rather than the level of the individual. Explain why.

Part C

In this remainder of this problem, we will re-analyze the Broockman and Kalla experiment by aggregating responses at the household level.

First, report the total number of **households** assigned to treatment and control respectively.

Next, evaluate whether we observe an effect of treatment assignment on contactability. Estimate the average treatment effect of a household being assigned to receive the transgender canvassing treatment on the probability that **at least one** member of the household was able to be contacted. Provide a 95% confidence interval and interpret your results. Is it reasonable to assume that drop-out due to non-contactability is as-good-as-random? Discuss why or why not.

Part D

Broockman and Kalla only elicit follow-up surveys from those respondents who were able to be contacted. Subset the data down to only those respondents who were able to be contacted and again, aggregate by household. This time, calculate whether at least one contacted person in a household responded to the follow-up surveys at t1, t2, t3, and t4. Estimate the average treatment effect of being assigned treatment versus placebo on the response probability at each of these follow-up points. Provide a 95% confidence interval for each and interpret your results. Do you find evidence for differential attrition between treated and control groups?

Part E

Subset the data down to only those respondents who were able to be contacted and who responded to the 3-day survey. Aggregate by household and calculate the household average feeling thermometer towards transgender individuals at the 3-day survey. Estimate the average treatment effect of a household receiving the transgender canvassing treatment on the 3-day feeling thermometer. Provide a 95% confidence interval and interpret your results. Would we conclude that there is a statistically significant ATE at $\alpha = .05$?

Part F

Repeat your analysis from Part E but using the **change** in the feeling thermometer from the baseline to the 3-day survey as the outcome. Compare the two sets of results and discuss any differences that you observe. Which approach is better and why?

Problem 3

In this problem, you will use simulation to learn about the sampling variance of the difference-in-means estimator for the ATE under different randomization schemes.

Assume the following data-generating process:

We observe a small sample of $N = 10$ observations. Each unit is assigned treatment $D_i = 1$ with some probability $Pr(D_i = 1)$. We will assume that the outcome is generated by $Y_i = \tau D_i + \epsilon_i$ where $\epsilon_i \sim \text{Normal}(0, 4)$. That is, ϵ_i is distributed *i.i.d* Normal with mean 0 and **variance** 4. We will assume a constant, additive treatment effect of $\tau = 4$ for the sake of the simulation.

Part A

Suppose treatment was assigned via independent Bernoulli trials with a constant probability of treatment $Pr(D_i = 1) = .5$ and $D_i \perp\!\!\!\perp D_j$ for all units $i \neq j$. Condition on there being at least one control group and one treated group respondent in the sample. Using a monte carlo simulation and assuming the data-generating process above, find the variance of the sampling distribution of the simple difference-in-means estimator for τ (use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations).

Part B

Now consider a completely randomized experiment where exactly $N_t = 5$ units receive treatment and exactly $N_c = 5$ units receive control. In this setting, the marginal probability of treatment is $\mathbb{P}(D_i = 1) = .5$ but D_i is not independent of D_j. Using a monte carlo simulation for this assignment process, find the variance of the sampling distribution of the simple difference-in-means estimator (again, use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations). Compare your variance to the variance under the data-generating process from Part A and discuss why they may differ.

Part C

Sometimes when designing an experiment, it is impossible to completely randomize over the entire sample of respondents since respondents arrive in a sequence. For example, experimenters fielding online surveys do not observe the entire sample and sometimes have to randomly assign treatments in a "just-in-time" manner.

Efron (1971) suggests an alternative approach to independent bernoulli randomization that biases the coin depending on how many units have previously been assigned to the treatment group versus the control group.

Consider the randomization scheme where treatment is assigned sequentially for units 1 through 10 according to their order. In other words, treatment for unit 1 is randomly assigned. Then treatment for unit 2 is randomly assigned depending on the value of the treatment for unit 1, and so on... Let $\tilde{N_{t,i}}$ denote the number units treated prior to unit i, $\tilde{N_{c,i}}$ the number of units under control prior to unit i and $\tilde{Z_i} = \tilde{N_{t,i}} - \tilde{N_{c,i}}$ or the difference in the number of treated and control groups. By definition, $\tilde{Z_1} = 0$ since there are no treated or control units when the first unit is assigned.

Define the probability of treatment $Pr(D_i = 1)$ for the ith unit as

$$Pr(D_i = 1) = \begin{cases} \pi & \text{if } \tilde{Z_i} < 0 \\ 0.5 & \text{if } \tilde{Z_i} = 0 \\ (1 - \pi) & \text{if } \tilde{Z_i} > 0 \end{cases}$$

Intuitively, the assignment mechanism biases the probability of receiving treatment upward if there are fewer treated than control and biases it downward if there are more treated than control at the time of assignment.

Let $\pi = .9$ and again, condition on there being at least one control group and one treated group respondent in the sample. Using a monte carlo simulation for this assignment scheme, find the variance of the sampling distribution of the simple difference-in-means estimator (use `60637` as your random seed set at the beginning of the code fragment and use 10000 monte carlo iterations). Compare your variance to your result in Part A and your result in Part B. Discuss any differences you observe.

Part D

Using your simulation results from Part C, is the difference-in-means estimator using this assignment scheme unbiased for the average treatment effect $\tau = 4$?

Part E

Intuitively, what will happen to the sampling variance if π is set to be less than .5? (You don't need to use a simulation to answer this, but you are welcome to use one if it would help).