MY457: Problem Set 1

Wed/29/Jan

This problem set is formative and will not contribute to your final grade. However, it is designed to build and deepen your conceptual understanding of the material and to practice applying the material in R. Using tools and resources such as ChatGPT and Stack Overflow is allowed and encouraged. Bear in mind that you are ultimately responsible for the work you submit, and that these tools often provide incorrect solutions. Make sure that however you use these tools it aligns with your best interests, and enhances your learning in this class.

This problem set must be submitted on Moodle by 12pm (noon) on Wed/05/Feb. You must also use the provided `.Rmd` template to produce a .pdf with your answers. If your submission is late, is not a .pdf, or is not appropriately formatted, you will not receive feedback on your work.

1 Concepts

This question reviews some of the concepts covered in class. Mathematical notation can be a useful tool to explain concepts, but it's important that you understand and can explain the concepts clearly and concisely. If you want to support your explanations with mathematical notation, this page provides a tutorial on including mathematical notation in Rmarkdown.

Consider a simple study of the effect of a treatment $D_i \in \{0, 1\}$ on Y_i for all $i \in \{1, 2, 3, \ldots, N\}$.

1.1. Explain the notation Y_{i1}. For the same unit i, when would this quantity be equal to Y_i?

1.2. What is the difference between $\mathbb{E}[Y_{i0}|D_i = 1]$ and $\mathbb{E}[Y_{i0}|D_i = 0]$? When would you expect these quantities to be equal? When would you expect them to be unequal?

1.3. Explain how randomly assigning individuals into treatment ($D = 1$) and control ($D = 0$) allows for the identification of the average treatment effect (ATE).

2 Simulations

In this question will use simulated data to test some of our intuitions about randomised experiments. The advantage of using a simulated dataset is that we have explicit control over the data generating process, and know the 'true' answer to any question we pose.

2.1. Suppose we are planning a trial to test the effectiveness of a policy on the wages of workers, with what we assume is a constant average treatment effect τ_{ATE}.

The real trial data that we collect will include information on participant characteristics, specifically age and education, along with actual treatment assignment (D), and the outcome variable (Y).

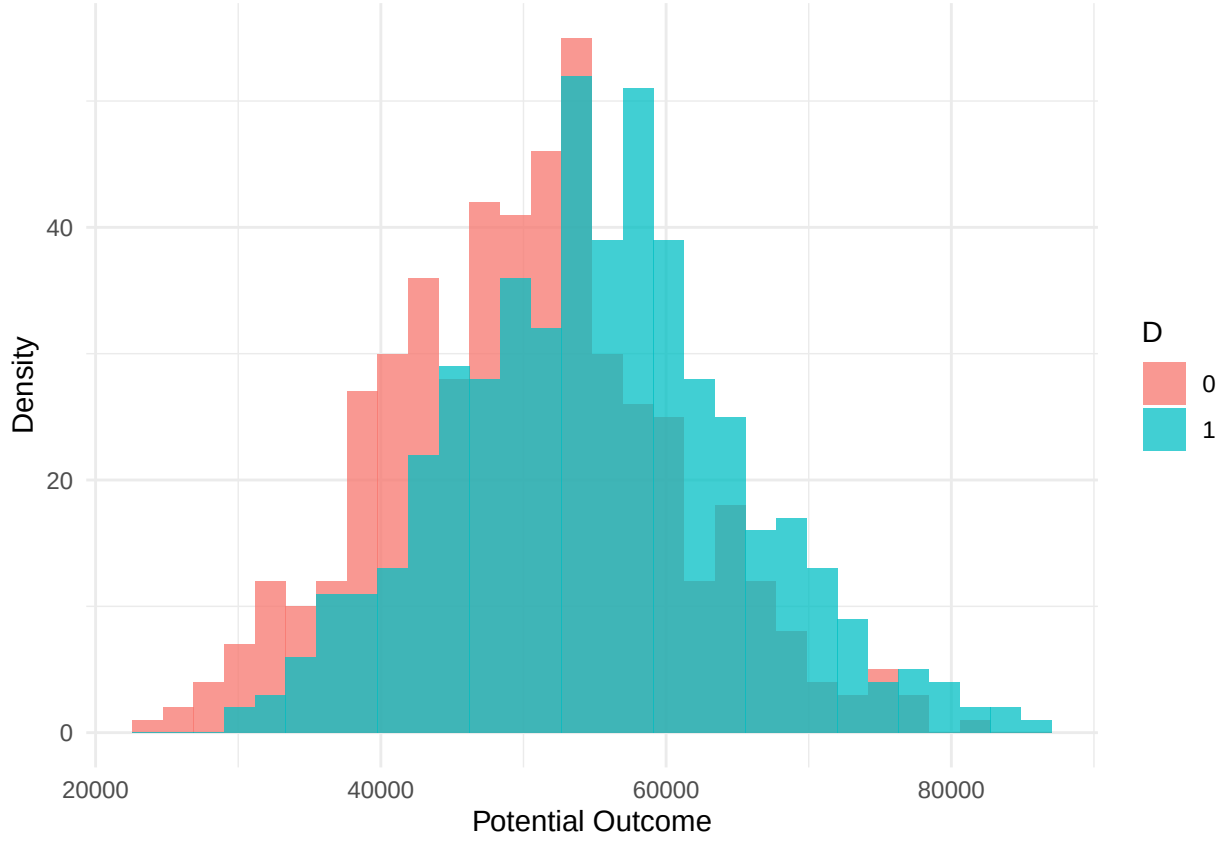
To explore some of the properties of our proposed trial, we first simulate some data. Explain in words what the code below does.

```
set.seed(123)

n  <- 500

tau <- 5000
```

```r
data <- data.frame(
  Age       = rnorm(n, mean = 42, sd = 10),
  Education = sample(1:4, n, replace = TRUE),
  Y0        = rnorm(n, mean = 50000, sd = 10000)
)

data <- data %>% mutate(
  Y1 = Y0 + tau,
  D  = sample(c(0, 1), n, replace = TRUE),
  Y  = ifelse(D == 1, Y1, Y0)
)
```

2.2. Randomisation implies the balancing principle. One way of testing for balance is by showing that the means of the two groups is statistically indistinguishable. Using a t-test, assess whether both age and education are balanced across conditions.

2.3. Let's look at the difference in **potential outcomes** by treatment condition. One attractive way of doing this is through a histogram of our potential outcomes, as shown below.



Calculate the difference between the potential outcomes as a difference-in-means (do not worry about statistical inference). What do you find? Is this surprising or unsurprising?

2.4. Let's assess the effect of our intervention on the **realised outcome** Y. First, generate a plot that shows the distribution of Y conditional on treatment status D. Here is a link to an introduction to ggplot for plotting. You can use any other package to generate this plot if you wish, including base `R`.

Second, estimate the average treatment effect (ATE). Since we have a randomized controlled trial, one

obvious estimator is the difference-in-means between the two groups. An alternative estimator would be linear regression (OLS). Implement both estimators. What do you notice? How does your result compare to the 'ground truth' answer you calculated in Question 2.3?

2.5. (Extra credit): Show that the answer to Question 2.4 was not due to chance (a 'lucky draw'). Over at least 1,000 repeat simulations (note: you will want to remove the `set.seed` step), generate a fresh draw of the data and estimate the sample-specific ATE with observed data. Calculate the difference between this and what you know to be the true (fixed) value of τ_{ATE} and store that difference. Finally, produce a histogram that shows the distribution of that difference over your repeated samples, along with its mean. What do you conclude?

3 Replication

In this question we will use real-world data to reinforce what we have learned. We will analyse the dataset employed in *How to Elect More Women: Gender and Candidate Success in a Field Experiment*.

In the USA, women tend to be underrepresented in legislative bodies. The authors designed a field experiment to test whether messages from party leaders can affect women's electoral success. The general belief is that there are two factors that explain why few women are elected. The first factor is related to the so-called 'supply side', with fewer female candidates vying for office. However, particularly among conservatives, voters' biases, also called the 'demand side', may play an important role. In the experiment messages are sent to leaders of precinct-level caucus meetings to see if tackling the demand side, the supply side, or both jointly, can increase the number of women who are elected.

The messages were divided into 4 categories: 1) Placebo control, 2) Supply messages, 3) Demand messages, and 4) Supply+Demand messages. In the data, group 1 is the control group receiving a placebo message unrelated to the aforementioned factors. Groups 2 to 4 represent the different treatment groups with messages relating to both factors.

3.1. Read into `R` the replication data file `how_to_elect_more_women.dta`. These data are at the precinct level. Using `prop_sd_fem2014`, the proportion of state delegates elected from the precinct in 2014 who were women, create a new dummy variable called `sd_onefem2014` that takes a value of 1 if at least one women was elected within the precinct in 2014, and 0 otherwise. This will be our outcome variable.

3.2. Show the proportion of precincts at each treatment/control group.

3.3. Take two pre-treatment variables of your choice that are not the outcome (`sd_onefem2014`) and test whether there is balance between the treatment groups.

3.4. Estimate the ATE of the different treatments. Hint: you can use three separate linear regressions, where you subset the data to just the control condition and one of the treatments, to estimate the effect of each treatment level. What do you find? Is there any effect of treatment. Why do you think this is happening?

3.5. (Extra credit): Simultaneously estimate the ATE of the different treatments using a single linear regression. Do any of your conclusions change?