Homework 1

January 12, 2022

Instructions

- This homework is due Wednesday, January 19 at 3pm EST.
- This week, please submit this homework via Canvas. We will talk in class about how to submit via GitHub in the future.
- Please format your homework solutions using R Markdown. You are welcome to simply add your answers below each question.
  - If the question requires a figure, make sure you have informative title, axis labels, and legend if needed.
  - Note: When I've given the framework of an answer's code, I've included the option `eval=FALSE` in the R chunk. When you start filling in your answer, you'll need to switch this to `eval=TRUE`.
- Turn in both the .rmd file and the knitted .pdf file.
  - Knitting the .rmd file to a .pdf file should help ensure your code runs without errors, but double check the .pdf output is what you expected.

Question 1 (Exercise 1 from GG Ch. 1)

In your own words, what is an experiment and how does it differ from an observational study?

Question 2 (Exercise 2 from GG Ch. 1)

Would you classify the study described in the following abstract as a field experiment, a naturally occuring experiment, a quasi-experiment, or none of the above? Why?

"This study seeks to estimate the health effects of sanitary drinking water among low-income villages in Guatemala. A random sample of all villages with fewer than 2,000 inhabitants was selected for analysis. Of the 250 villages sampled, 110 were found to have unsanitary drinking water. In these 110 villages, infant mortality rates were, on average, 25 deaths per 1,000 live births, as compared to 5 deaths per 1,000 live births in the 140 villages with sanitary drinking water. Unsanitary drinking water appears to be a major contributor to infant mortality."

Question 3

Using DeclareDesign, declare a population with 100 people. Include age as a covariate. Assume `age` in our population is distributed $\texttt{age} \sim N(40, 10)$.

Next, imagine (i.e., declare) potential outcomes for each person in your sample. We are interested in the outcome of the person's salary. If a person is treated they attend a job training program, and if a person is untreated they do not attend the program. *If* everyone attended the job training program, assume the average salary for the population would be 65K. *If* everyone did *not* attend the job training program, assume the average salary for the population would be 55K. Also assume some heterogeneity in the population's salaries whether they attend the job training program or not. We'll assume a similar amount of heterogeneity in both cases, so in either condition, add noise distributed $N(0, 5K)$.

Finally, create your `design` object by adding these two declarations together.

```
library(DeclareDesign)

population <- # declare population here

potential_outcomes <- # declare potential outcomes here

design <- population + potential_outcomes
```

Question 4

Using the `draw_data` function, draw one hypothetical population's potential outcomes using the provided code. Print the first ten lines of your dataset. Then, answer the following questions.

```
single_draw <- draw_data(design)
single_draw[1:10, ]
```

4a

Create a column for the causal effect on unit i. Call it `tau`.

4b

Create three histograms. One for the potential outcomes if treated, one for the potential outcomes if not treated, and one for the individual causal effects of treatment. Comment on each.

4c

Describe in words the fundamental problem of causal inference in terms of the columns in our dataframe `single_draw`.

4d

Calculate the ATE.

4e

Can we observe the ATE?

Question 5

Now, instead of examining *one* hypothetical study like we did in Question 4, let's simulate many studies. Specifically, draw 1000 hypothetical studies, calculate the ATE for each, and plot a histogram of the ATE's. Comment on what you see in the histogram. (Note: there are many ways to accomplish this in code.)