Problem Set 1

Instructions

If you work with any other person or persons, you must list their names and describe in detail who worked on what and how. Every individual who turns in a problem set is responsible for knowing how to do every single answer. Assume that it is always possible that I will give you a different version of the problem if I suspect plagiarism.

Citing sources, especially for code solutions, is acceptable and encouraged.

Causal Inference Conceptual Questions

Question 1

a. Explain the notation "$Y_i(0)$" and "$Y_i(1)$."

b. Explain and contrast the notation "$Y_i(0)|D_i = 1$" with the notation "$Y_i(0)|d_i = 1$."

c. Explain and contrast the meaning of "$E[Y_i(0)]$" with the meaning "$E[Y_i(0)|D_i = 1]$."

d. Explain why $E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] = 0$ when D_i is randomly assigned.

Question 2

Assume that we conduct a randomized experiment. Define the ATT as $E[\tau_i|D_i = 1]$. Below is the proof of the following statement, "When units are randomly assignment to treatment and control, the ATT is equal to the ATE in expectation. Equivalently, $E[\tau_i|D_i = 1] = E[\tau_i]$." Explain substantively what each line in the following proof does and why. For notation purposes, the first m observations are the treatment group, and the remaining N-m observations are the control group.

1. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[\frac{\sum_i^m Y_i}{m}] - E[\frac{\sum_{m+1}^N Y_i}{N-m}]$
2. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = \frac{E[Y_1] + ... + E[Y_m]}{m} - \frac{E[Y_{m+1}] + ... + E[Y_N]}{N-m}$
3. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[Y_i(1)|D_i = 1] - E[U_i(0)|D_i = 0]$
4. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = E[Y_i(1)] - E[Y_i(0)]$
5. $E[\frac{\sum_i^m Y_i}{m} - \frac{\sum_{m+1}^N Y_i}{N-m}] = ATE$

Question 3

a. What is a standard error? What is the difference between a standard error and a standard deviation?

b. How is randomization inference used to test the sharp null hypothesis of no effect for any unit?

c. Define a 95% confidence interval.

d. How does complete random assignment differ from block random assignment and cluster random assignment?

e. Designs that assign the same number of units to treatment and controlled are known as "balanced designs." Explain at least one desirable statistical property of these designs.

Question 4

Assume that the treatment effect is constant τ. Show that the $V[Y_i(0)] = V[Y_i(1)]$ and show that $\rho[Y_i(0), Y_i(1)] = 1$.

Question 5

Consider the abstract from Andrews, Delton, and Kline (2021) below and answer the first three questions we ask about every research design as if you were the authors:

"Disaster responses are political. But can citizens make useful disaster decisions? Potential obstacles are that such decisions are complex, involve public goods, and often affect other people. Theories of political decision-making disagree on whether these problems can be overcome. We used experimental economic games that simulate disaster to test whether people are willing and able to prevent disasters for others. Groups of players face a complex task in which options that might help vary in their riskiness. Importantly, although all options are reasonable, which option is most useful depends on the experimental condition. We find that players will pay to help, can identify which option is most useful across experimental conditions, and will pay to learn how best to help. Thus, players were able to make useful and costly decisions to prevent others from experiencing disaster. This suggests that, in at least some situations, citizens may be able to make good disaster decisions."

Question 6

Based on experience, your instructor has noticed that students who study at least 3 hours a day are more likely to get good scores on exams. I, therefore, recommend in my syllabus that students study for my class for at least 3 hours a day.

a. What is the implied treatment (D_i) of the underlying experiment?

b. State the potential outcome under control and the potential outcome under treatment for the underlying experiment in words.

c. Since your instructor noticed this pattern, we can consider this a design in which Oski assigned a particular realization of d_i to each student. Why might you be skeptical of the causal effect I'm claiming?

R Conceptual Questions

Question 1

Sliver is running a special on its pizza of the day. Customers can get two 12" pizzas for the price of one 18". Is this a cost-effective deal for the consumer in terms of the amount of pizza for the price? Use R to justify your answer.

Question 2

Using R, determine how many of the following five numbers are divisible by 2 or 7: 2, 9, 14, 21, 69

Consider using the `ifelse` function. The easiest way to solve this problem is to use the modulo operator. In R, the modulo operator is `%%`.

Question 3

(Taken from Project Euler)

If we list all the natural numbers below ten that are multiples of 3 or 5, we get 3,5,6, and 9. The sum of these multiples is 23.

Find the sum of all the multiples of 3 or 5 below 1000.

Applied Questions

Question 1

Oski needs some help working at the Botanical Garden. He asks six first-year students to donate either 30 minutes or 60 minutes to improve the space. Due to time constraints, he wants to decide on a procedure to allocate slots such that an equal number of students is in the 30 minute and 60-minute groups. Oski considers two ways to do this.

Method 1: Flip a fair coin for each student. If heads, assign them to the 30-minute group. If tails, assign them to the 60-minute group.

Method 2: Put three slips of paper with 30 minutes written on them and three slips of paper with 60 minutes written on them in a box. For each student, draw a slip of paper without replacement and assign the student to the number written.

a. Using R (setting the randomization seed to 2356) show why the first method has a problem that the second method does not.

b. Using R, change the number of subjects to 600 instead of six, adjusting the number of slips of paper in the box appropriately. Does anything change with Method 1? Which method do you prefer as N gets large?

Question 2

A study program advertises that all students who participate will improve their grades more than they otherwise would without the program. Ten students participate in the program. Compared to a normalized baseline of 0, the scores for the treatment group are $\{1, 0, 0, 4, 3\}$ and the scores for the control group are $\{-5, 4, 7, -7, -4\}$.

State the sharp null hypothesis for this design. Using R (set your randomization start seed to 1234567), show the results of a randomization test of the sharp null hypothesis of no effect of the study program with a balanced design. Do you believe the program's claims of success?