**P8122 - STATISTICAL METHODS FOR CAUSAL INFERENCE**

**MIDTERM**

**DUE OCTOBER 23RD 11:59PM**

**INSTRUCTIONS**

* Upload a pdf of your MIDTERM EXAM on Canvas along with your R Script on the due date.
* The R script should be self-contained, so someone else should be able to run it as is and get your results.
* The R script should also be well-commented, so it is clear which code goes with which question.
* You ARE NOT ALLOWED to discuss these problems with each other nor share code with each other. This must be your own work.
* Please be concise!

**QUESTION 1**

**(80 points)**

Consider the following table of 20 individuals and suppose you know all the potential outcomes. In this table:

A=1: assigned new treatment

A=0: assigned standard treatment

Y=1: disease prevented

Y=0: disease not prevented

|  |  |  |
| --- | --- | --- |
| **“Truth”** | | |
| **Individual** | **Y1** | **Y0** |
| 1 | 1 | 0 |
| 2 | 0 | 0 |
| 3 | 1 | 0 |
| 4 | 0 | 0 |
| 5 | 1 | 1 |
| 6 | 0 | 0 |
| 7 | 1 | 0 |
| 8 | 0 | 0 |
| 9 | 1 | 0 |
| 10 | 0 | 0 |
| 11 | 0 | 1 |
| 12 | 0 | 0 |
| 13 | 0 | 0 |
| 14 | 1 | 1 |
| 15 | 1 | 0 |
| 16 | 0 | 1 |
| 17 | 1 | 0 |
| 18 | 1 | 0 |
| 19 | 0 | 0 |
| 20 | 1 | 0 |

1. (4 points) Define and calculate the appropriate measure for the effect of the treatment and interpret it.

Suppose you conduct a study in the real world on the same 20 individuals to estimate the effect of the new treatment vs. the standard treatment on preventing disease. Here are the data from **Study 1**.

|  |  |  |  |
| --- | --- | --- | --- |
| **Study 1** | | | |
| **Individual** | **A** | **Y|A=1** | **Y|A=0** |
| 1 | 1 | 1 | . |
| 2 | 1 | 0 | . |
| 3 | 0 | . | 0 |
| 4 | 0 | . | 0 |
| 5 | 0 | . | 1 |
| 6 | 0 | . | 0 |
| 7 | 1 | 1 | . |
| 8 | 0 | . | 0 |
| 9 | 0 | . | 0 |
| 10 | 0 | . | 0 |
| 11 | 1 | 0 | . |
| 12 | 1 | 0 | . |
| 13 | 1 | 0 | . |
| 14 | 0 | . | 1 |
| 15 | 0 | . | 0 |
| 16 | 1 | 0 | . |
| 17 | 1 | 1 | . |
| 18 | 1 | 1 | . |
| 19 | 1 | 0 | . |
| 20 | 0 | . | 0 |

1. (4 points) Write the formula for the appropriate measure for this study, calculate it and interpret it.
2. (3 points) Compare this estimate with what you obtained from Question 1. If you obtain the same number, explain why. If you obtain a different number, explain why (hint: make sure your explanation discusses the assignment mechanism).
3. (4 points) Explain how this type of data might arise in (a) an observational study and (b) a randomized controlled trial.
4. (2 points) Can you rule out a particular study design given what you observe?

Now let’s suppose you want to perform a study (**Study 2**) that is a randomized controlled trial (RCT) looking at the effect of the new treatment vs. the standard treatment on preventing disease. You perform this study on the 20 individuals from Question 1. Assuming you have knowledge of every individuals’ potential outcomes under both treatments (i.e., you know the truth), how would you create an RCT in which the assignment mechanism was randomized using a block randomization method, ensuring it is probabilistic and unconfounded.

1. (6 points) Please describe your process in words and provide a table showing your study (i.e., your observed data) under this assignment mechanism.
2. (6 points) Write and test the sharp null hypothesis of no causal effect in your study (describe your process and plot the randomization distribution). Interpret your results.
3. (6 points) Provide the point estimate and confidence interval inverting the hypothesis testing procedure you conducted in (7). Describe the procedure and interpret your results.
4. (6 points) Provide the point estimate and confidence interval of the marginal average causal effect in your study using Neyman’s approach. Interpret your results.
5. (6 points) Compare the estimates you obtained in (8) and (9) with what you computed in **(1)**.

Let’s say you also want to perform an observational study looking at the effect of the new treatment vs. the standard treatment on preventing disease (**Study 3**). You perform this study on 40 new individuals sampled from the population. It is hypothesized that the new treatment is better for disease prevention than the standard treatment. It is also hypothesized that individuals with normal white blood cell (WBC) counts (L=1) are more likely to be prescribed the new treatment and also more likely to have a better disease prognosis (i.e., more likely to have disease prevented) compared with individuals with abnormal WBC counts (L=0).

1. (3 points) What is your scientific question of interest for this observational study under Neyman’s approach to inference?
2. (3 points) Represent the situation described in a DAG. Be sure to label and define the variables you include in the DAG.
3. (3 points) What does the DAG imply about the crude association? Based on the DAG and your hypothesis drew, leaving L unadjusted for in your analysis would lead to which of the following situations. Circle your choice and explain the reasoning for your choice.
   1. No bias in the estimated effect
   2. Bias in the estimate such that it overestimates the average causal effect (i.e., the estimate is larger in magnitude than the true causal effect).
   3. Bias in the estimate such that it underestimates the average causal effect (i.e., the estimate is smaller in magnitude than the true causal effect).

Explanation for choice:

Below is the data from **Study 3**, the observational study of 40 new individuals sampled from the population.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study 3** | | | | |
| **Individual** | **Y|A=1** | **Y|A=0** | **A=a** | **L=l** |
| 1 | 1 | . | 1 | 1 |
| 2 | . | 0 | 0 | 1 |
| 3 | . | 0 | 0 | 0 |
| 4 | . | 0 | 0 | 0 |
| 5 | 1 | . | 1 | 1 |
| 6 | 0 | . | 1 | 0 |
| 7 | . | 0 | 0 | 1 |
| 8 | . | 0 | 0 | 0 |
| 9 | . | 0 | 0 | 0 |
| 10 | 0 | . | 1 | 1 |
| 11 | . | 1 | 0 | 0 |
| 12 | 1 | . | 1 | 1 |
| 13 | 1 | . | 1 | 1 |
| 14 | . | 1 | 0 | 0 |
| 15 | . | 1 | 0 | 0 |
| 16 | 1 | . | 1 | 1 |
| 17 | 1 | . | 1 | 0 |
| 18 | 1 | . | 1 | 0 |
| 19 | 1 | . | 1 | 0 |
| 20 | 1 | . | 1 | 1 |
| 21 | . | 1 | 0 | 1 |
| 22 | 1 | . | 1 | 0 |
| 23 | 0 | . | 1 | 0 |
| 24 | 0 | . | 1 | 0 |
| 25 | . | 1 | 0 | 0 |
| 26 | . | 0 | 0 | 0 |
| 27 | . | 0 | 0 | 0 |
| 28 | 1 | . | 1 | 0 |
| 29 | . | 0 | 0 | 0 |
| 30 | 0 | . | 1 | 0 |
| 31 | . | 1 | 0 | 1 |
| 32 | 1 | . | 1 | 1 |
| 33 | . | 0 | 0 | 0 |
| 34 | . | 1 | 0 | 0 |
| 35 | 1 | . | 1 | 1 |
| 36 | 1 | . | 1 | 1 |
| 37 | . | 1 | 0 | 1 |
| 38 | 1 | . | 1 | 0 |
| 39 | . | 0 | 0 | 0 |
| 40 | . | 1 | 0 | 0 |

1. (6 points) Compute the estimate and confidence interval for ACE using the g-formula for observational studies. Interpret the result.
2. (3 points) Compare your result with what you obtained in **Question 1**. If you get approximately the same result, explain why. If you get something different, explain why.
3. (3 points) Do you find support for your hypothesis in the data? Please explain why or why not and, where applicable, show any additional calculations you performed to reach your conclusion.

In truth, the relationship between A and Y is much more complex. See below the true DAG where A is the treatment, Y is the outcome, L, H, F and B are measured variables and U is unmeasured.

Chart, line chart

Description automatically generated

1. (4 points) Identify a variable or set of variables in the DAG that when conditioned on would close all back-door paths between A and Y.
2. (2 points) What is the relationship between NUCA assumption of potential outcomes and a DAG?
3. (3 points) Identify a variable in the DAG that when conditioned on would open a closed path from A to Y.
4. (3 points) Conceptually, what is a collider and why is it problematic to adjust for a collider? Can you provide an example of a collider in the DAG?

**Question 2**

**(20 points)**

One of our group readings focused on the benefits of diversity in health care. This question is inspired by this paper.

Suppose that the Department of Health of New York State surveys hospitals to find out how many doctors from minority backgrounds are employed and have leadership positions (e.g., attending/head of divisions). The Department learns that some hospitals have been hiring a striking majority (>95%) of white doctors for leadership positions. The Department therefore creates a workshop for hospital administrators that focuses on the benefits of diversity in leadership. The workshop is given at all hospitals that currently have majority of white doctors in leadership positions. At other hospital across the state, the workshop is given if requested by hospital administrators. Two years later, the Department again asks all hospitals to indicate the number of doctors from minority backgrounds that were promoted to leadership positions.

1. Identify the units, potential outcomes, treatment, and any observed covariates. Explain each

in one sentence.

2. Define in words and with a mathematical formula the causal effect you would be interested in studying.

3. How would you describe the study design in terms of the assignment mechanism? Is this study design appropriate to address your question? Explain your reasoning.

4. If you were to advise the New York State Department of Health, how would you suggest them to proceed (from design to analysis) in order to quantify the causal effect of interest?