Causal Inference - Homework 1

Introduction to Causal Inference 097400
Spring 2025

Submission date: May 6th, 2025

Question 1

Using potential outcomes notation, provide the following:

I. Give an example of a data generating process (DGP, i.e. a joint distribution) which includes a binary covariate X, a binary treatment T and two potential outcomes Y_0 and Y_1.

II. For the DGP you created, calculate the expected potential outcomes $\mathbb{E}[Y_t]$ for $t \in \{0, 1\}$ and the resulting ATE.

III. Let $Y = TY_1 + (1 - T)Y_0$. Give an example of a DGP whose $\hat{ATE}$ depends on if we assume $(Y_1, Y_0) \perp\!\!\!\perp T$ or $(Y_1, Y_0) \perp\!\!\!\perp T|X$. What is the ATE?
This can be the same DGP as I, but you may not copy the example from the tutorials.

IV. Calculate the $\hat{ATE}$ under each assumption. Show your calculations for $\hat{\mathbb{E}}[Y|T = 1], \hat{\mathbb{E}}[Y|T = 0], \hat{\mathbb{E}}[Y|T = 1, X = 1], \hat{\mathbb{E}}[Y|T = 0, X = 1], \hat{\mathbb{E}}[Y|T = 0, X = 1]$ and $\hat{\mathbb{E}}[Y|T = 0, X = 0]$.

Please define the DGP by filling out a table of the following form:

Index	X	T	Y_0	Y_1	Y
1					
2					
3					
4					
5					
6					

Question 2

I. Let the following variables be defined: X covariates, T binary treatment, Y the outcome, and the propensity score $e_t(x) = \mathbb{E}[\mathbb{I}\{T = t\}|X = x] = P(T = t|X = x)$. We define

$$q_t(x) = \frac{\mathbb{I}\{T = t\}}{e_t(x)}$$

to be a signed re-weighting function. Under the assumptions of ignorability, consistency, and positivity prove that:

$$\mathbb{E}[q_t(x)Y] = \mathbb{E}[Y_t].$$

II. Imagine that a drug is tested in a Randomized Control Trial (RCT) which does not include any pregnant woman. Would $q_t(x)y$ be defined? What does that mean for our ability to infer the effect the drug would have on a pregnant woman?

Question 3

(adapted from a problem by Dr. Daniel Nevo)

In a recent study, researchers wished to estimate the effect of receiving a basketball as a gift for the 10th birthday ($T \in \{0, 1\}$) on whether the child was accepted to college with a full scholarship ($Y \in \{0, 1\}$).
You have received a data file `Data.csv`, which contains information of 8 participants from 4 different families. For each participant, you have his potential outcomes under each combination of treatment assignments, given to all the participants.
The treatment vector is given as a concatenated string, in the format of $t_0t_1 \cdots t_8$, where t_i indicates whether participant i received the gift. For example, the column "$Y00000001$" represents the potential outcome in the case where only participant 8 has received such a gift.

1. In the given data, does the SUTVA assumption hold? back your answer with example.

2. Assume that the SUTVA assumption holds *between families*. Under this assumption, define the new potential outcomes and treatments. Hint: The potential outcomes can be reduced to a new single treatment variable with 3 levels.

3. Calculate the 3 possible average treatment effects (i.e. compare each two levels of the new treatment).

It us not necessary to submit your code.

Question 4

Give two examples of real-world data with features X, treatment T and one or more observed outcome variables Y.
For each example:

1. Formulate a causal question, i.e. a treatment variable and an outcome. Explicitly define the treatment and potential outcomes.

2. State whether you believe there is confounding between the treatment variable and the outcome. Explain briefly.

3. Give an example of a prediction problem related to the data that does not require causal reasoning.

Examples can come from the fields of politics, biology, sports, economics, entertainment, medicine, transportation and so on - use your imagination. You might find the article " *A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks*" by Miguel Hernán (available on the course Moodle) to be helpful. Do not use examples from Hernán's article.