Exercise 1: Introduction to Causal Inference

Question 1: Consider the following statement from Mayer-Schonberger and Cukier (2013) "Big Data: A Revolution That Will Transform How we Live, Work, and Think", page 14:

"Correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening. And in many situations this is good enough. If millions of electronic medical records reveal that cancer sufferers who take a certain combination of aspirin and orange juice see their disease go into remission, then the exact cause for the improvement in health may be less important than the fact that they lived. ... we can let the data speak for itself."

Other than the fact that they mix up singular and plural by stating that we should let the "data" (plural) speak for "itself" (singular) :-), describe in one paragraph (no longer than one half page) why this statement is problematic. Provide an example illustrating how their interpretation of the scenario may be erroneous.

Question 2) In randomized controlled trial settings, researchers are often interested in estimating *per protocol effects*. Consider a simple scenario with a randomization indicator R, with $R = 0$ denoting "assigned to placebo" and $R = 1$ denoting "assigned to treated", an adherence indicator A, with $A = 0$ denoting "did not adhere" and $A = 1$ denoting "adhered by taking treatment on the day randomized", and an outcome variable Y, with $Y = 1$ denoting "event", and $Y = 0$ denoting "no event". Can you write the per protocol effect, defined as being assigned to treatment and adhering relative to being assigned to placebo and adhering, using potential outcomes notation? Write these effects on the risk difference, risk ratio, and odds ratio scales.

Question 3) Suppose we conduct a study of the the effect of 6 mg Dexamethasone daily versus placebo on a measure of lung function one week after admission to the hospital due to respiratory symptoms resulting from infection with SARS-CoV-2. Suppose we let Y denote lung function at the end of seven days, and D_j denote Dexamethasone treatment on day j of follow-up (e.g., $D_j = 1$ denotes treated with Dexamethasone on day j; $D_j = 0$ denotes not treated with Dexamethasone on day j). Please describe, in words, the effect that the following contrast of potential outcomes captures:

$$\psi = E(Y^{d_1=1,d_2=1,d_3=1,d_4=1,d_5=0,d_6=0,d_7=0}) - E(Y^{d_1=1,d_2=1,d_3=1,d_4=0,d_5=0,d_6=0,d_7=0})$$

Question 4) Please re-write the right-hand side of the equation in Question 3 more compactly (instead of writing out the exposure value on each of the seven days).

Question 5): Please complete the Table under SUTVA:

ID	Exposure (A)	Outcome (Y)	Y(a=1)	Y(a=0)
1	1	1		
2	1	1		
3	0	1		
4	1	0		
5	0	0		
6	0	1		

Question 6): Suppose we are interested in the effect of quitting smoking on high blood pressure. Do you think the average treatment effect or the effect of treatment on the treated is more relevant? Explain why or why not.

Question 7): Again, for the example of the relation between quitting smoking and high blood pressure, can you describe a scenario where we may collect some data and where the no interference assumption would be violated?

Question 8): Consider the following statement from a paper by Athey et al (2020)[https://arxiv.org/pdf/1909.02210.pdf], page 14: In the setting of interest we have data on an outcome Y_i, a set of pretreatment variables X_i and a binary treatment $W_i \in \{0, 1\}$. We postulate that there exists for each unit in the population two potential outcomes $Y_i(0)$ and $Y_i(1)$, with the observed outcome equal to corresponding to the potential outcome for the treatment received, $Y_i = Y_i(W_i)$.

What assumption(s) are the authors relying on when they say "We postulate that there exists . . . "? Why?

Question 9): Consider the exchangeability assumption. Why is the word "exchangeable" used to describe this concept? What, precisely, is being exchanged?

Question 10): Consider a regression model with an exposure and 11 confounders, for a total of 12 variables:

$$E(Y \mid X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \ldots + \beta_{12} C_{11}$$

What is the total number of possible interactions in this model? What are the total number of 2-way

interactions? Show your reasoning.

Question 11): Suppose you had superpowers and were able to measure potential outcomes. Suppose you used these measures to fit a model that regresses the exposure A against all measured confounders C (i.e., propensity score model), and that there was no measured confounding, selection bias, and information bias (i.e., exchageability was met). If you included the potential outcomes in the regression model:

$$\text{logit}\{P(A = 1 \mid C, Y^a)\} = \beta_0 + \beta_1 C_1 + \ldots + \beta_p C_p + \theta Y^a$$

Can you determine from this information alone what the value of θ is if exchangeability holds? Can you determine what the value of θ is if exchangeability doesn't hold?

Question 12) Consider a two-arm placebo controlled randomized trial with four mutually exclusive strata labeled $S = 1, S = 2, S = 3$ and $S = 4$. Suppose that the treatment was assigned to: 20% of individuals in stratum $S = 1$; 30% of individuals in stratum $S = 2$; 15% of individuals in stratum $S = 3$; and 10% of individuals in stratum $S = 4$. Can you determine all of the propensity score values in the sample of individuals in the trial?

(Bonus?) Question 13) Using the information provided in Question 12, please write a logistic regression equation that the defines the propensity score for this randomized trial. What are the parameter values in this logistic regression model?