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.

The major problem with the above statement is that it is encouraging a complete disregard for the possibility of selection bias, information bias, or confounding bias playing a role in the detected correlation. Specifically, each of these, alone or in combination, can be a potential explanation for the fact that a correlation is detected between aspirin + orange juice, and cancer remission. For example, the true causal effect of aspirin + orange juice on cancer remission may be null. But a correlation may be detected if, for example, (*i*) individuals taking a particular cancer drug are more likely to take aspirin (due to drug side effects, such as headache) and are more likely to drink orange juice (again, due to some effect of the drug, which may result in citrus cravings); and (*ii*) the drug causes cancer remission. In this case, telling patients that they should consume more aspirin and orange juice may do nothing to help them, and may actually result in additional problems (e.g., side effects of aspirin). Bottom line: DATA DON'T SPEAK.

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.

We can define per protocol effects on the risk difference, risk ratio, and odds ratio scales using potential outcomes as follows:

$$RD = P(Y^{a=r=1} = 1) - P(Y^{a=1,r=0} = 1)$$

$$RR = P(Y^{a=r=1} = 1)/P(Y^{a=1,r=0} = 1)$$

$$OR = \frac{P(Y^{a=r=1} = 1)}{P(Y^{a=r=1} = 0)} \bigg/ \frac{P(Y^{a=1,r=0} = 1)}{P(Y^{a=1,r=0} = 0)}$$

where $P(Y^{a,r} = 1)$ is the probability of the outcome that would be observed if adherence A were set to some value a, and randomization arm R were set to some value r.

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:

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

This effect captures the expected difference in lung function that would be observed if all individuals took dexamethasone each day until (and including) day 4, and then no dexamethasone thereafter, relative to taking dexamethasone each day until (and including) day 3, and then no dexamethasone thereafter.

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).

$$\psi = E(Y^{\overline{d}_4=1,\underline{d}_5=0}) - E(Y^{\overline{d}_3=1,\underline{d}_4=0})$$

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		

The key to filling out this table is to recognize that, under SUTVA, only one of the potential outcomes for each individual is identifiable (the outcome under the observed exposure). Thus, for each row, one of the potential outcomes should be left blank (the potential outcome under the exposure status that was not observed).

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.

While the average treatment effect is informative in this scenario, it is difficult to conceive of a population where everyone smokes versus where everyone doesn't smoke. Consequently, it is difficult to conceive of a population where everyone has the opportunity to quit or not quit smoking. For this reason, one can easily argue that the ETT is more relevant in this setting, since it measures the effect of quitting smoking among those who actually quit.

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?

Smoking has a number of "second hand" effects. For example, if a husband is a smoker, and a wife is not, the fact that the husband smokes may affect the wife's blood pressure via second hand smoke effects. Consequently, the no interference assumption in this scenario may be violated if data are collected from individuals in close proximity to one-another. In this case, one person's outcome may be affected by the other person's exposure status, which is a violation of no interference.

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?

The authors are implicitly invoking SUTVA, which consists of the combination of counterfactual consistency and no interference. Mathematically, the statement $Y_i = Y_i(W_i)$ states that the observed outcome is the potential outcome under the observed exposure. This is the definition of counterfactual consistency. However, in writing this, they are also implicitly stating that individual i's outcome is only dependent upon individual i's exposure, and not others' exposure. For this reason, the second sentence in the above invokes both counterfactual consistency and no interference.

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

The key problem we are addressing with identification assumptions are to overcome the fundamental problem of causal inference. This problem describes the issue that we cannot observe individuals in both their exposed and unexposed state. As a result, we cannot directly compute average outcomes that would be observed if everyone were exposed and unexposed.

When exchangeability holds, we are able to assume that the average outcome that was observed among the unexposed can stand in for the missing risk that would have been observed had exposed individuals been unexposed. Similarly, we can assume that the average outcome that was observed among the exposed can stand in for the missing risk that would have been observed had unexposed individuals been exposed. In effect, the missing risks are "exchangeable" between the exposed and unexposed.

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.

This is simply an application of the relevant combinatorial equations. For the total number of possible interactions, which includes 2-way, 3-way, all the way up to 12-way, we use the multiplication rule:

$$2^{12} - 12 - 1 = 4,083$$

There are a total of 4,083 interactions possible with 12 variables.

For the total number of 2-way interactions, this is an application of choosing 2 out of 12:

$$\binom{12}{2} = \frac{12!}{2!(12-2)!} = 66$$

There are a total of 66 two way interactions.

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?

Exchangeability implies independence between the exposure and the potential outcomes. Thus, if exchangeability holds, one can infer that the value of θ will be zero. Alternatively, if exchangeability is violated, one can only infer that the value of θ will not be zero.

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?

The propensity score is defined as the probability of receiving the treatment. In the context of this example, the only possible values the propensity score can take are .2, .3. .15, and .1.

(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?

In the scenario described, the four propensity score values (and thus, four strata) can each be represented by a single parameter in a logistic regression model. Equivalently, one can describe the four mutually exclusive strata using three dummy variables. Thus, the propensity score model can be written as:

$$P(Y = 1 \mid S) = \text{expit}\{\beta_0 + \beta_1 I(S = 2) + \beta_2 I(S = 3) + \beta_3 I(S = 4)\}$$

where β_0 represents the stratum for the referent group, in this case $S = 1$, and where the coefficients for $I(S = 2)$, $I(S = 3)$, and $I(S = 4)$ represent the *difference* between the baseline coefficient value and the other strata values

One can also compute the specific values for the β's needed to make this logistic regression model compatible with the scenario described. First, note that the S groups are mutually exclusive. Meaning that when, e.g., $S = 1$, then $I(S = 2) = I(S = 3) = I(S = 4) = 0$ (and similarly for when S takes on other values). Second, note that the value for β_0 should return a predicted probability of 0.2, since this represents the PS for those in stratum 1. Thus, we can simply use the inverse of the expit function (i.e., the logit function) to compute the value for β_0 as:

$$\text{logit}\{0.2\} = \beta_0 \implies \log(0.2/0.8) = \beta_0 \implies \beta_0 = -1.39$$

Next, we work to solve for β_1, which should yield a probability of 0.3, the propensity score for those in stratum $S = 2$. This time, however, we have to account for the presence of the referent parameter, the intercept.

$$\text{logit}\{0.3\} = -1.39 + \beta_1 \implies \log(0.3/0.7) + 1.39 = \beta_1 \implies \beta_1 = 0.539$$

Solving for β_2 and β_3 can be done in exactly the same way as we solved for β_1.

$$\text{logit}\{0.15\} = -1.39 + \beta_2 \implies \log(0.15/0.85) + 1.39 = \beta_2 \implies \beta_2 = -0.348$$

$$\text{logit}\{0.1\} = -1.39 + \beta_3 \implies \log(0.1/0.9) + 1.39 = \beta_3 \implies \beta_3 = -0.812$$

We can easily use R to check our work:

```
expit <- function(x){1/(1+exp(-x))}
ps_mod <- as.matrix(c(-1.39,.539,-.348,-.812))


S1 <- c(1,0,0,0)
S2 <- c(1,1,0,0)
S3 <- c(1,0,1,0)
S4 <- c(1,0,0,1)


round(expit(S1%*%ps_mod),2)
```

```
##      [,1]
## [1,]  0.2
```

```
round(expit(S2%*%ps_mod),2)
```

```
##      [,1]
## [1,]  0.3
```

```
round(expit(S3%*%ps_mod),2)
```

```
##      [,1]
## [1,] 0.15
```

```
round(expit(S4%*%ps_mod),2)
```

```
##      [,1]
## [1,]  0.1
```