Discussion Assignment 2

Introduction to Causal Inference (Biostat683)

Assigned: October 4, 2021
Group presentation: In class on October 18, 2021. Group slides must be uploaded to GoogleDrive by 2pm on Oct 18. Please make sure to include your names and group number on the slides. One presentation per group
Individual write-ups: NO individual-level write-ups required. BUT you must work with your group on the slides and make them as detailed as possible. If 2+ group members disagree on a given question, you can include multiple slides to represent differing responses and their justifications.
Study: If you were assigned Study 1 in assignment #1, then you will complete answers for Study 2 in this assignment. Likewise, if you were assigned Study 2 in assignment #1, then you will complete answers for Study 1 in this assignment.

1 Instructions

For the previous two studies, **think** through the following questions. **Using the following background (Steps 0-2 of the Causal Roadmap), please provide brief *written* answers to the questions for your assigned study.** Use the notation developed in class. You are encouraged to discuss as a group, but please submit your own written responses. **As a group, be prepared to *present* your answers in class** (October 18, 2021).

2 Background (Steps 0-2 of the roadmap) for Studies #1-2

2.1 Study #1: Physical activity and mortality in the elderly

- **Step 0:** What is the effect of free-living energy expenditure on the seven-year survival among older active adults.
- **Step 1:** Consider the following structural causal model $\mathcal{M}^*$. Other causal models from assignment 1 may also be correct, but to simplify discussion we will work with the following.
  - Endogenous variables: $X = (W, A, Y)$
  - Exogenous variables: $U = (U_W, U_A, U_Y) \sim \mathbb{P}_U$
  - Structural equations F:

$$W = f_W(U_W)$$
$$A = f_A(W, U_A)$$
$$Y = f_Y(W, A, U_Y)$$

    where W={smoking, comorbidities, body fat}, A={energy expenditure} and Y={survival}. By collapsing the baseline variables into a single node W, we lose the information that smoking was measured first. However, there are no consequences for our statistical estimation problem.
  - There are no independence assumptions on $\mathbb{P}_U$.
- **Step 2:** Suppose we are interested in the average treatment effect (i.e., the causal risk difference).
  - The target causal parameter is the difference in the counterfactual probability of survival if all elderly adults had a high energy expenditure ($A = 1$) and the counterfactual probability of survival if all elderly adults had a low energy expenditure ($A = 0$)

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$

where the counterfactual outcome Y_a is the seven-year survival for an individual if, possibly contrary to fact, they had energy expenditure $A = a$ for this two week period.

2.2 Study #2: Effect of male circumcision on risk of HIV acquisition

- **Step 0:** What is the effect of male circumcision on HIV acquisition after two years? More specifically, what is the causal effect of male circumcision on the probability of becoming infected with HIV over two years in this rural Kenyan population?
- **Step 1:** Consider the following structural causal model $\mathcal{M}^*$. Other causal models from assignment 1 may also be correct, but to simplify discussion we will work with the following.
  - Endogenous variables: $X = (W, A, Z, Y)$
  - Exogenous variables: $U = (U_W, U_A, U_Z, U_Y) \sim \mathbb{P}_U$
  - Structural equations F:

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Z &= f_Z(W, A, U_Z) \\ Y &= f_Y(W, A, Z, U_Y) \end{aligned}$$

    where W={tribe, religion}, A=male circumcision, Z={sexual behavior, STI}, and Y=HIV status
  - There are no independence assumptions on $\mathbb{P}_U$.
- **Step 2:** Suppose we are interested in the average treatment effect (i.e., the causal risk difference).
  - The target causal parameter is the difference in the counterfactual risk of HIV acquisition if all males were circumcised ($A = 1$) and the counterfactual risk of HIV acquisition if all males were not circumcised ($A = 0$)

$$\Psi^*(\mathbb{P}^*) = \mathbb{E}^*(Y_1) - \mathbb{E}^*(Y_0)$$

    where the counterfactual outcome Y_a is the two-year HIV status for an individual, if possibly contrary to fact, they had circumcision status $A = a$.

3 Questions to be answered

1. *Step 3: Observed data & link to causal model.*
   (a) Specify the observed data.
   (b) What notation do we use to refer to the distribution of the observed data?
   (c) What is the link between the structural causal model (SCM) and the observed data?
   (d) What is the statistical model $\mathcal{M}$? Does the SCM place any restrictions on $\mathcal{M}$?
2. *Steps 4-5: Identifiability & Committing to a statistical estimand*
   (a) Using the backdoor criterion, assess identifiability of $\Psi^*(\mathbb{P}^*)$.
   (b) If not identified, under what assumptions would it be? Are some of these sets of additional assumptions more plausible than others? Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?
   (c) Specify the target parameter of the observed data distribution (i.e., the statistical estimand).
   (d) What is the relevant positivity assumption? Are you concerned about violations of the positivity assumption in your study?
3. *Study-specific question for Study 1:* Suppose the investigators assume no unmeasured common causes of (W, A, Y) is this necessary for identifiability? Is it sufficient?
4. *Study-specific question for Study 2:*The study investigators adjust for (condition on) W={tribe, religion} and Z={sexual behavior, STI}. Under what causal structure would W would satisfy the back door criterion, but (W, Z) would not? Under what causal structure would (W, Z) satisfy the back door criterion, but W alone would not?