R Lab One

Katherine Wolf
Introduction to Causal Inference (PH252D)
February 29, 2020

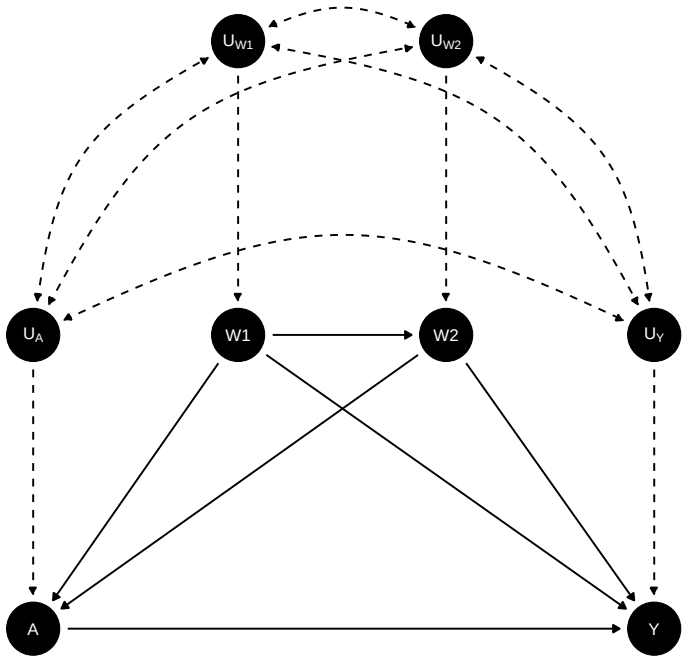
1 Background Story

2 Steps 1-5 of the Roadmap

2.1 Step 1: Causal model representing real knowledge

a. Draw the accompanying directed acyclic graph (DAG).

The directed acyclic graph is



where the endogenous nodes $\boldsymbol{X} = (W1, W2, A, Y)$ include

- $W1$, a binary indicator variable representing potable water access, where $W1 = 1$ if the child had access to potable water at study initiation and $W1 = 0$ otherwise;
- $W2$, a binary indicator variable representing prior infectious disease within the two weeks prior to the study initiation, where $W2 = 1$ if the child suffered from an infectious disease within the two weeks prior to the study initiation and $W2 = 0$ otherwise;
- A, a binary indicator variable representing the exposure of interest, ready-to-use therapeutic food (RUTF), where $A = 1$ if the child received RUTF and $A = 0$ if the child received the standard supplement; and

- Y, a continuous variable representing the outcome of interest, the child's weight gain in pounds at study termination;

each of the background exogenous variables $\boldsymbol{U} = (U_{W1}, U_{W2}, U_A, U_Y) \sim \mathbb{P}_U$ represents all the unmeasured factors for the $\boldsymbol{X}$ variable denoted in its subscript that determine the values that that variable in $\boldsymbol{X}$ takes; and the structural equations F are

- $W1 = f_{W1}(U_{W1})$
- $W2 = f_{W2}(W1, U_{W2})$
- $A = f_A(W1, W2, U_A)$
- $Y = f_Y(W1, W2, A, U_Y)$

b. Are there any exclusion restrictions? Recall we are working with recursive (time-ordered) structural causal models.

Aside from the inherent exclusion restrictions in a time-ordered structural causal model (wherein each variable is excluded from the parent sets of the variables before it in the order, on the assumption that it happened after the events represented by the preceding variables and thus could not have affected them), no.

c. Are there any independence assumptions on the distribution of unmeasured factors $\mathbb{P}_U$?

No, not from this story. We do not know, for example, if the intervention represented by A was randomly assigned (and thus if we can assume that its unmeasured factors don't exist or are independent of the unmeasured factors influencing the values of the other variables in $\boldsymbol{X}$).

2.2 Step 2: Counterfactuals and causal parameter

a. Define the counterfactual outcomes of interest with formal notation and in words.

The counterfactual outcomes of interest are, formally,

$(Y_a : a \in A) \sim P_{U,X}$, where $A = \{0, 1\}$, or, equivalently,

$Y_a = f_Y(W1, W2, a, U_Y)$ for $a \in A = \{0, 1\}$.

In words, the counterfactual outcome Y_a is the weight gain in pounds of an individual child at study termination if, possibly contrary to fact, that child had intervention (i.e., supplement) status $A = a$. Thus

- Y_1 is the counterfactual weight gain in pounds of an individual child at study termination if the child received the RUTF supplement, and
- Y_0 is the counterfactual weight gain in pounds of an individual child at study termination if the child received the standard non-RUTF supplement.

b. How are counterfactuals derived?

Counterfactuals are derived by intervening on the structural causal model (SCM) M^F to set $A = a$. The counterfactual distributions are implied by those joint distributions of $(\boldsymbol{U}, \boldsymbol{X})$ that contained in the set of allowed counterfactual distributions that the SCM M^F defines.

c. Suppose we are interested in the average treatment effect. Specify the target causal parameter. Use formal notation as well as explain in words.

The average treatment effect is the difference between the expected (here, mean) weight gain for children in the population at study termination if all of them had received the RUTF supplement and the expected weight gain for children in the population at study termination if all of the children in the population had received the standard supplement.

Formally, the average treatment effect is

$\Psi^F(P_{U,X}) = E_{\boldsymbol{U},\boldsymbol{X}} Y_1 - E_{\boldsymbol{U},\boldsymbol{X}} Y_0 = E_{\boldsymbol{U},\boldsymbol{X}}(Y_1 - Y_0)$,

where

- $\Psi^F(P_{U,X})$ is the average treatment effect Ψ as a function of the structural equations F,
- $E_{\boldsymbol{U},\boldsymbol{X}}Y_1$ is the expected weight gain for children in the population at study termination if everyone in the population had received the RUTF supplement, i.e., , i.e., the expected value of Y if A were set to $a = 1$ for everyone, and
- $E_{\boldsymbol{U},\boldsymbol{X}}Y_0$ is the expected weight gain for children in the population at study termination if everyone in the population had received the standard supplement, i.e., the expected value of Y if A were set to $a = 0$ for everyone.

2.3 Step 3: Observed data and link to causal

a. Specify the link between the SCM and the observed data.

We assume that the observed data were generated by sampling from a data-generating system compatible with the SCM. In other words, we posit a set of variables that represent the observed data in the random variable $\boldsymbol{O} \sim P_O$, which are a subset of the endogenous variables $\boldsymbol{X}$, where the distribution of the exogenous variables $\boldsymbol{U}$, P_U and the structural equations $\boldsymbol{F}$ identify the distribution of $\boldsymbol{O}$. Here, $\boldsymbol{O} = \boldsymbol{X}$, so $P_O(O = o) = \sum_n P_f(X = x|U = u)P(U = u) = \sum_n I(X(u) = x)P(U = u)$

b. What restrictions, if any, does the SCM place on the allowed distributions for the observed data? (Recall d-separation.)

The SCM places no restrictions on the allowed distributions for P_O, the distribution of the observed data, because none of the nodes are d-separable, and thus we cannot assume that any variable in $\boldsymbol{O}$ is independent of any other variable $\boldsymbol{O}$.

c. What notation do we use to denote the true (but unknown) distribution of the observed data and the statistical model?

P_O denotes the true but unknown distribution of the observed data, and $\mathbb{M}$ denotes the set of possible distributions for the observed data, also known as the statistical model, where $P_O \in \mathbb{M}$.

2.4 Steps 4-5: Identification and statistical estimand

a. Using the backdoor criterion, assess identifiability.

A set of variables W satisfies the backdoor criterion with respect to (A, Y) if no node in W is a descendant of A and W blocks all paths from A to Y that include an arrow *into* A. No subset of the observed covariates W here satisfies the backdoor criterion because no means exists either to block the open path from A to Y through the unobserved variables U_A, U_W1, U_W2, and U_Y, and thus the target causal parameter $\Psi^F(P_{U,X}) = E[Y_1 - Y_0]$ is not identified by any parameter $\Psi(P_O)$ in the observed data distribution.

b. If the target causal parameter is not identified, under what assumptions would it be?

Any of the following sets of assumptions would allow identification of the target causal parameter:

- A is independent of Y, $W1$, and $W2$. Then $W1$ would satisfy the backdoor criterion, and adjustment for it would identify the target causal parameter.
- Y is independent of A, $W1$, and $W2$. Then $W2$ would satisfy the backdoor criterion, and adjustment for it would identify the target causal parameter.

Since investigators can generally manipulate the independence of the exposure A more easily than the outcome Y, my answers to the following questions use the first set of assumptions.

c. What notation is used to denote the original SCM augmented with additional assumptions needed for identifiability?

$\mathbb{M}^{\mathbb{F}*}$.

d. Specify the target parameter of the observed data distribution (the statistical estimand).

Assuming that A is independent of Y, $W1$, and $W2$, under $\mathbb{M}^{\mathbb{F}*}$:

$\Psi(P_O) = \sum_{w1} E_O(Y|A = a, W1 = w1)P_O(W1 = w1) = E_{O,W1}[E_O(Y|A = 1, W1) - E_O(Y|A = 0, W1)]$

e. What is the relevant positivity assumption? Is it reasonable here?

The relevant positivity assumption is that each combination of values for the exposure A and the potable water access covariate $W1$ must occur with some positive probability. Formally, $\min_{a \in \mathbb{A}} P_O(A = a|W1 = w1) > 0$ for all $w1$ for which $P_O(W1 = w1) > 0$. As this assumption is about the observed data, we can check it using the observed data. And, indeed, children exist in all four possible category combinations for the exposure of RUTF supplement or standard supplement and the covariate representing access to potable water or no access to potable water. Thus the assumption is reasonable here. (make table?)

3 Bonus: Identifying the Mean Outcome Under a Dynamic Intervention

1. **Explain why (1) holds using properties of conditional expectations. Given access to the full population and the ability to implement intervention *d*, what does (1) tell you about how you could compute $\mathbb{E}_{U,X}[Y_d]$?**
2. **Explain why (2) holds using properties of conditional expectations and the fact that $Y_d \perp\!\!\!\perp A|W_1, W_2$ under our convenience assumptions for the backdoor criterion made in Question 4 of Section 2.**
3. **Explain why (3) holds. What does this mean in terms of the RUTF example?**
4. **Explain why (4) holds. What does this mean in terms of the RUTF example?**

4 A Specific Data-Generating Process

4.1 Closed form evaluation on the target parameter

1. **Evaluate the target causal parameter $\psi^F(\mathbb{P}_{U,X})$ in closed form (i.e., by hand) for this data generating process.**

   $\mathbb{E}_{U,X}(Y_a) = \mathbb{E}_{U,X}[4 * a + 0.7 * W1 - 2 * a * W2 + U_Y]$

   $\Psi^F(\mathbb{P}_{U,X}) = E_{O,W1}[E_O(Y|A = 1, W1) - E_O(Y|A = 0, W1)] =$

   $\mathbb{E}_{U,X}[4 * 1 + 0.7 * W1 - 2 * 1 * W2 + U_Y] - \mathbb{E}_{U,X}[4 * 0 + 0.7 * W1 - 2 * 0 * W2 + U_Y] =$

```
overall_expectation <- 3.2 - 0.4*(exp(0.5) / (1 + exp(0.5)))

# 2.951016
```

2. **Interpret $\psi^F(\mathbb{P}_{U,X})$.**

   The counterfactual expected weight gain at study termination for children in this population would be 2.951 pounds higher if all children received the RUTF supplement than if all children received the standard supplement.

4.2 Translating this data generating process for $\mathbb{P}_{U,X}$ into simulations, generating counterfactual outcomes and evaluating the target causal parameter.

1. **First set the seed to 252.**

```
set.seed(252)
```

2. **Set $n = 50,000$ as the number of independent and identically distributed draws from the data-generating process.**

```
n = 50000
```

3. **Simulate the background factors U.**

```
U_W1 <- runif(n, min = 0, max = 1)
U_W2 <- runif(n, min = 0, max = 1)
U_A <- runif(n, min = 0, max = 1)
U_Y <- rnorm(n, mean = 0, sd = 0.3)
```

4. **Evaluate the structural equations F to deterministically generate the endogenous nodes X.**

```
plogis(1)

## [1] 0.7310586
```

5. **Intervene to set the supplement to RUTF $(A = 1)$ and generate counterfactual outcomes Y_1 for n units. Then intervene to set the supplement to the standard $(A = 0)$ and generate counterfactual outcomes Y_0 for n units.**

6. **Create a data frame X to hold the values of the endogenous factors (W_1, W_2, A, Y) and the counterfactual outcomes Y_1 and Y_0. The rows are the n children and the columns are their characteristics. Use the head and summary to examine the resulting data.**

7. **Evaluate the causal parameter $\psi^F(\mathbb{P}_{U,X})$ for this population of 50,000 units.**

5 Defining the Target Causal Parameter with a Working Marginal Structural Model

1. **For $n = 5,000$ children, generate the exogenous factors U and the pre-intervention covariates $(V, W1, W2)$. Then set $A = 1$ to generate the counterfactual weight gain under RUTF Y_1. Likewise, set $A = 0$ to generate the counterfactual weight gain under the standard supplement Y_0.**
2. **Create a data frame `X.msm` consisting of age V, the set treatment levels a, and the corresponding outcomes Y_a.**
3. **Evaluate the target causal parameter.**
4. **Interpret the results.**