Discussion Assignment #2

Sabrina Boyce, Shelley Facente, and Steph Holm

10/23/2019

Question 2(d) What are the observed data O? Factorize the observed data distribution P_0 according to the time-ordering.

N i.i.d. copies of $O = (L_0, A_0, L_1, A_1, L_2) \sim P_0$

(with Y as a subset of the Ls)

Basic probability reminder for factoring:

- $P(A, B) = P(A \cup B) = P(A|B) \times P(B)$
- If we know that say 10% of the class drinks coffee, and that of coffee drinkers 50% like ginger cookies, then 5% of the class drinks coffee AND likes ginger cookies.
- So then, $P(A, B, C) = P(A|B, C) \times P(B|C) \times P(C)$

Because we have something that is a function of the joint distribution (P_0) we can factorize it using probability rules into conditional distributions that we can use.

Question 2(d) What are the observed data O? Factorize the observed data distribution P_0 according to the time-ordering.

$$
\begin{aligned}
&P(L_0, A_0, L_1, A_1, L_2) = \\
&P(L_2 | A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times \\
&P(A_1 | L_1 = l_1, A_0 = a_0, L_0 = l_0) \times P(L_1 | A_0 = a_0, L_0 = l_0) \times \\
&P(A_0 | L_0 = l_0) \times P(L_0 = l_0)
\end{aligned}
$$

$$
\begin{aligned}
&P(L_0, A_0, L_1, A_1, L_2) = \\
&\qquad \prod_0^2 P(L_k = l_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1}) \times \\
&\qquad \prod_0^1 P(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1})
\end{aligned}
$$

Question 2(d) What are the observed data O? Factorize the observed data distribution P_0 according to the time-ordering.

$$P(O = o) =$$

$$\prod_{0}^{2} Q(L_k = l_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1}) \times$$

$$\prod_{0}^{1} g(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1})$$

Question 2(e): What assumptions are needed to identify the causal parameter from the observed data distribution?

In order to identify our causal parameter from the observed data distribution, we must assume:

1. Sequential Randomization, which assumes each a is independent of y (and which allows us to control for a different set of covariates at each time point)
2. Positivity, which assumes that no strata have few or no observations.
3. No unmeasured confounding
4. Correct model is specified
5. No measurement error

Question 2(f): Specify the G-computation formula (i.e. the statistical estimand) for this example.

The general longitudinal G computation formula is:

$$E(Y_{\bar{a}}) = \sum_{\bar{l}} E(Y | \bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K) \quad \times$$
$$\prod_{t=1}^{K} P(L_t = l_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$$

Therefore in this specific example it would be:

$$E(Y_{\bar{a}}) = \sum_{\bar{l}} E(Y | \bar{A}_2 = \bar{a}_2, \bar{L}_2 = \bar{l}_2) \quad \times$$
$$\prod_{0}^{2} P(L_t = l_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$$