MATH-449 - Biostatistics
EPFL, Spring 2022
Problem Set 2

1. (Modified version of an exercise from Vanessa Didelez) Suppose we investigate the effect of a specific motivational training program on the performance of students in elementary school. We recruit several school classes, and in each class we randomly allocate one half of the students to the motivational training, while the other half serves as the control group.

   a) It is likely that the students in the motivation group share what they learned in the motivational training with the other students. Which of the following assumptions is/are violated: exchangeability, positivity or consistency? Why? **Consistency says that if $A = 0$, then Y will equal $Y^{a=0}$. Here, we are told that some students with $A = 0$ in these observed data receive exposure to essential ingredients of $A = 1$ (the information taught in the motivational training). It is then plausible that for these students with $A = 0$, Y will equal $Y^{a=1}$.**

   b) Name sources of (unobserved) confounding in this study, if any. **Treatment is randomly assigned and patients comply with their nominal exposure group, so there will be no causes of exposure except randomization, among trial participants.**

2. Prove the following Lemma:

   Lemma 1. *If V follows a NPSEM-IE, then for any $p(\overline{v}_{j-1})$ with $p(\overline{v}_{j-1}) > 0$ we have that $p(v_j \mid \overline{v}_{j-1}) = p(v_j \mid pa_j)$ and therefore the joint density factorizes as*

$$p(v) = \prod_{j=1}^{m} p(v_j \mid pa_j).$$

   Proof. First consider a factorization of $p(v)$ that uses only laws of probability:

$$p(v) = \sum_{\epsilon_1,\ldots,\epsilon_m} \prod_{j=1}^{m} p(v_j \mid \overline{v}_{j-1}, \epsilon) \prod_{k=1}^{m} p(\epsilon_k \mid \overline{\epsilon}_{k-1}).$$

   By the fact that V_j is a deterministic function of its parents and its error term, following the definition of the NPSEM:

$$p(v) = \sum_{\epsilon_1,\ldots,\epsilon_m} \prod_{j=1}^{m} I(v_j = f_{v_j}(pa_j, \epsilon_j)) \prod_{k=1}^{m} p(\epsilon_k \mid \overline{\epsilon}_{k-1}).$$

   By the definition of the NPSEM-IE, the epsilons are mutually independent, so:

$$p(v) = \sum_{\epsilon_1,\ldots,\epsilon_m} \prod_{j=1}^{m} I(v_j = f_{v_j}(pa_j, \epsilon_j)) \prod_{k=1}^{m} p(\epsilon_k).$$

   Re-arranging terms and using laws of probability:

$$p(v) = \prod_{j=1}^{m} \Big\{ \sum_{\epsilon_j} I(v_j = f_{v_j}(pa_j, \epsilon_j)) p(\epsilon_j) \Big\}$$

   Note that V_1 is a deterministic function of ϵ_1. Since V_2 is maximally a function of V_1 and ϵ_2, then V_2 is a deterministic function of $\overline{\epsilon}_2$. Arguing by forward induction on k, assume that for each $j = 1, \dots, k-1$, V_j is a deterministic function of $\overline{\epsilon}_j$. V_k is a deterministic function of ϵ_k and $\overline{V}_{k-1}$ by the NPSEM. By the inductive hypothesis, then $\overline{V}_{k-1}$ is a deterministic function of $\overline{\epsilon}_{k-1}$, so V_k is a deterministic function of $\overline{\epsilon}_k$. Then it follows by the NPSEM-**IE** that for all j, ϵ_j is independent of $\overline{V}_{j-1}$, and specifically, $PA_j \subseteq \overline{V}_{j-1}$. Thus we can conclude the proof by writing:

$$\begin{aligned}
p(v) =& \prod_{j=1}^{m} \Big\{ \sum_{\epsilon_j} I(v_j = f_{v_j}(pa_j, \epsilon_j)) p(\epsilon_j \mid pa_j) \Big\} \\
=& \prod_{j=1}^{m} \Big\{ \sum_{\epsilon_j} p(v_j \mid pa_j, \epsilon_j)) p(\epsilon_j \mid pa_j) \Big\} \\
=& \prod_{j=1}^{m} p(v_j \mid pa_j).
\end{aligned}$$

□

3. a) Prove the following equivalence: $\mathbb{E}[\mathbb{E}[Y \mid A = a, L]] = E[Y \frac{I(A=a)}{P(A=a|L)}]$, thus establishing the equivalence between the classical formulation of the g-formula, and its inverse-probability-weighted (IPW) representation.

   b) Is the IPW parameter a function of the propensities $P(A = a \mid L)$? Explain your answer.

Proof.

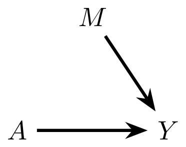
$$\begin{aligned}
\mathbb{E}[\mathbb{E}[Y \mid A = a, L]] =& \mathbb{E}[\mathbb{E}[Y \frac{I(A = a)}{P(A = a \mid L)} \mid A, L]] \\
= \mathbb{E}[Y \frac{I(A = a)}{P(A = a \mid L)}].
\end{aligned}$$

□

Previously we showed that the g-formula $\mathbb{E}[\mathbb{E}[Y \mid A = a, L]]$ is not a function of the propensities. Since the IPW expression is simply a re-formulation of the g-formula, then it too is not a function of the propensities (i.e. it is parameter that is variationally independent of the propensities, so long as the g-formula is well-defined - that is the positivity condition holds.

4. Consider a study of individuals who were randomly assigned to an exercise program ($A = 1$) or no exercise program ($A = 0$). Suppose the individuals were only allowed a daily intake of 2500 KCAL (calories) every day during the study period (define daily calories consumption as a mediator M). The individuals were carefully monitored such that everybody adhered to the dietary protocol. At the end of the study, Body Mass Index (BMI) was measured in each individual (BMI is the outcome Y). Draw a causal DAG depecting the data generating mechanism that produced the observed data. Use counterfactual notation to express the causal effect of the exercise program on BMI, when daily intake of calories is fixed to 2500 KCAL. Which assumptions are needed to identify this effect? Do these assumptions hold in the study described above?

   The estimand is $\mathbb{E}[Y^{a=1,m=2500} - Y^{a=0,m=2500}]$. In this trial, we need to observe patients who had $M = 2500$ for both treatment groups (positivity), and that $Y^{a,m} \perp\!\!\!\perp A, M$ for $a \in \{0, 1\}$ and $m = 2500$, which is guaranteed by randomization. However, we also need that $Y = Y^{a,m=2500}$ whenever $A = a$ and $M = 2500$. But this will likely depend on the type of calories (which foods, etc.), so consistency will likely fail:



one person who had $A = 1, M = 2200$ but got their calories through burgers and fries may have the same outcome as if they had $A = 1, M = 2800$ but got their calories through a plant-based diet.

5. Prove the following graphoid axiom (intersection):

$$\text{if } p(x, y, z, w) > 0 \text{ then } X \perp\!\!\!\perp W \mid Y, Z \text{ and } X \perp\!\!\!\perp Y \mid W, Z \implies X \perp\!\!\!\perp Y, W \mid Z$$

Proof. By the premised independencies, we have the following equalities, for some functions f, g, $\tilde{f}$, $\tilde{g}$:

$$\begin{aligned} p(x, y, z, w) =& f(x, y, z)g(y, w, z) \\ p(x, y, z, w) =& \tilde{f}(x, w, z)\tilde{g}(y, w, z) \end{aligned}$$

.

Taking equalities and re-arranging terms yields the following (where the positivity condition ensures the expression is well-defined):

$$\begin{aligned} f(x, y, z) =& \tilde{f}(x, w, z)\frac{\tilde{g}(y, w, z)}{g(y, w, z)} \\ =& \tilde{f}(x, w, z)g^*(y, w, z). \end{aligned}$$

Then note that the left hand side is not a function of w, so:

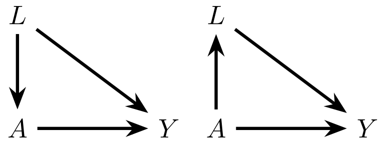
$$f(x, y, z) = f^\dagger(x, z)g^\dagger(y, z).$$

Plugging in to the first expression we then have:

$$p(x, y, z, w) = f^\dagger(x, z)g(y, w, z)g^\dagger(y, z),$$

From which it is easy to see then that $X \perp\!\!\!\perp Y, W \mid Z$.

□



6. Consider the graphs above. Can we use observed data on L, A, Y to assess whether either of these graphs above describe the true data generating mechanism (that is, the true causal model).

No - each of the graphs puts the same restrictions (none) on the observed data distribution, so we can not distinguish between these graphs without additional knowledge.