Causal Inference for Health Data (Winter 2026)
STATS C160/C260
Homework 1

Exercise 1. Basic probabilities

Seventy percent of cancer cases in a certain population are diagnosed in an early stage. Of those diagnosed early, 60% of the patients went to routine consultations twice a year, whereas 90% of the patients diagnosed late did not.

(a) Suppose a certain person has developed cancer and goes to routine consultations. What is the probability that cancer will be diagnosed early?

(b) Construct a probability distribution over three random variables X, Y, Z such that $(X \perp\!\!\!\perp Y)$ but $(X \perp\!\!\!\perp Y \mid Z)$ does not hold. You can either describe the full joint distribution or their conditionals.

Exercise 2. Estimation and Independence Relations

Consider random variables X_1, X_2, and Y and assume our goal is to compute the query $Q = P(y \mid x_1, x_2)$. We do not have any prior information about the conditional independence relations among these variables.

(a) For every one of the following sets of distributions, show how to compute the query Q based on them, or explain why this is not possible:

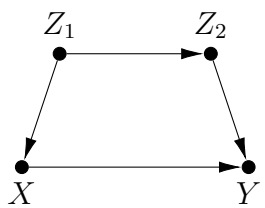
1. $P(x_1, x_2), P(y), P(x_1 \mid y)$, and $P(x_2 \mid y)$
2. $P(x_1, x_2), P(y)$, and $P(x_1, x_2 \mid y)$
3. $P(x_1 \mid y), P(x_2 \mid y)$, and $P(y)$
4. $P(x_1), P(x_2)$, and $P(x_1, x_2 \mid y)$
5. $P(x_1), P(x_2), P(x_1 \mid y)$, and $P(x_2 \mid y)$

(b) Suppose we learned that $(X_1 \perp\!\!\!\perp X_2 \mid Y)$ holds in P. Now, which of the sets before are sufficient to compute the query Q? Show how or explain why it is not possible.

Exercise 3. Query Estimation

Consider the following graphical model $\mathcal{G}$ below:

Suppose we want to compute the query $Q = \sum_{Z_1} P(y \mid x, z_1)P(z_1)$.



(a) Is $Q = P(y \mid x)$ in $\mathcal{G}$? Justify your answer.

(b) Suppose we have access to the marginal distribution $P(X, Y, Z_2)$. Is it possible to estimate Q? If so, show how to do it. Otherwise, explain why that is not the case.

Exercise 4. Specifying Structural Causal Models

The following is a description of a clinical decision support (CDS) alert system used in a hospital:

- The delivery channel location of the alert (L) could be via the clinician's *phone* or via the *EHR system* and is chosen at random every time with equal likelihood.
- Based on the clinician's age group (A), either a *text-only* alert or a *text+image* alert is used as the alert modality (M). The age group is either '*less than 40*' or '*40 or more*'.
- The clinician will see (S) the alert with the following probabilities:
  1. If $L =$ *phone* with probability 1/3,
  2. if $L =$ *EHR system*, $A =$'*less than 40*' with probability 1/5, and
  3. if $L =$ *EHR system*, $A =$'*40 or more*' with probability 1/6.
- The clinician will judge the alert to be clinically relevant (I) 60% of the time if the modality is *text+image*, or 40% of the time if the modality is *text-only*.
- The clinician has permission to administer the drug (D) in 40% of cases.
- The clinician will administer the drug (C) if the alert is seen, judged relevant, and the clinician has permission to administer the drug.

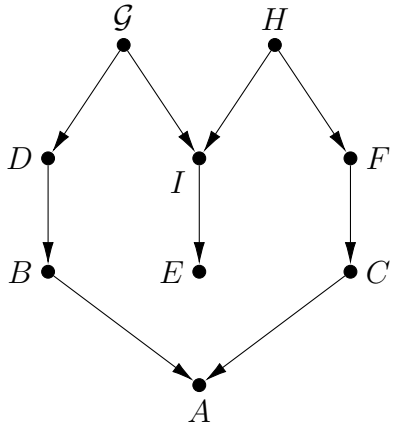
Variables L, S, M, I and C are observable, while A and D are unobservable (other unobservables, which are not mentioned, may be present as well).

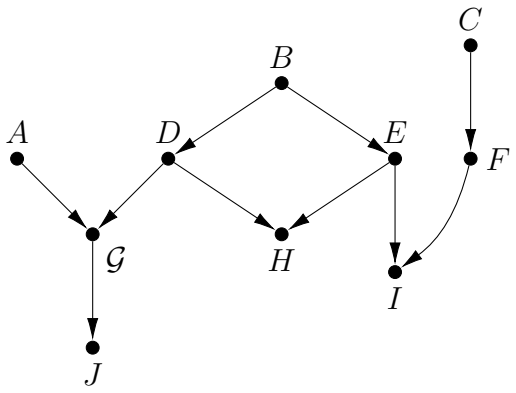
(a) Specify a structural causal model $\mathcal{M} = \langle \mathbf{V}, \mathbf{U}, \mathcal{F}, P(\mathbf{U}) \rangle$ that captures this setting. Make a reasonable choice for the distribution of A and the way M is determined.

(b) Draw the causal diagram corresponding to the given SCM.

Exercise 5. d-Connectedness

Consider the two graphs below, $\mathcal{G}$ (left) and $\mathcal{G}'$ (right).



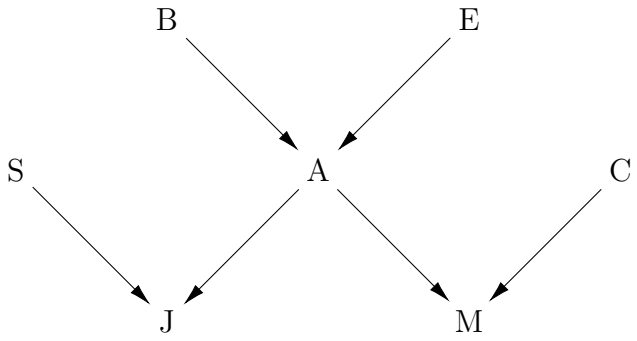


(a) List the variables that are d-connected to A given $\{B\}$ in $\mathcal{G}$.

(b) List the variables that are d-connected to A given $\{J\}$ in $\mathcal{G}'$.

Exercise 6. d-Separation

Consider the following graphical model,



and the conditional probability tables:

P(B=1)	P(E=1)	P(S=1)	P(C=1)
0.02	0.01	0.80	0.15

B	E	$P(A=1\|BE)$	A	C	$P(M=1\|AC)$	A	S	$P(J=1\|AS)$
0	0	0.01	0	0	0.05	0	0	0
0	1	0.3	0	1	0	0	1	0
1	0	0.9	1	0	0.85	1	0	0.97
1	1	0.98	1	1	0.15	1	1	0.1

(a) List all d-separation statements that hold assuming that $J = 1$.

(b) Compute the given probabilities using the graph and the probability tables:

(i) $P(M = 1)$

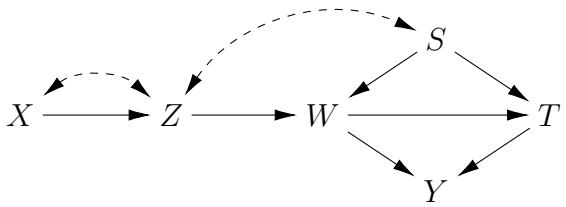
(ii) $P(J = 1|C = 0)$

(iii) $P(E = 1|M = 1, B = 0)$

(iv) $P(M = 1|B = 1, J = 0)$

Exercise 7. d-Separating Sets

Let $\mathcal{G}$ (shown below) be the causal diagram of some unknown model $\mathcal{M}$ and let P be $\mathcal{M}$'s observational distribution.



(a) Find a minimal set $\mathbf{A}$ (if it exists) that d-separates X and W.

$\mathbf{A} = \{\square\}$

(b) Find a minimal set $\mathbf{A}$ (if it exists) that d-separates X and S.

$\mathbf{A} = \{\square\}$

(c) Find **all** minimal set $\mathbf{A}$ (if any exists) that d-separate Z and Y.

$\mathbf{A} = \left\{ \boxed{}, \boxed{}, \boxed{}, \boxed{} \right\}$

Note: Not necessarily 4 minimal sets.

(d) [Harder] Draw a graph $\mathcal{G}'$ over the variables $\{X, S, W, Y\}$ such that

- $\mathcal{G}'$ has exactly the same independence as $\mathcal{G}$ with respect to $P(X, S, W, Y)$, and
- $\mathcal{G}'$ has the minimum number of edges while satisfying the constraint in the previous bullet.

Hint: In the first step, determine which graph is obtained if Z and T are unobserved, and list the independences implied by this graph. In the second step, consider which of the edges can be removed without affecting any of the independences.

Exercise 8. [Optional] d-Separation Theory

Let $\mathcal{G}$ be a causal diagram (may include bidirected arrows) over a set of variables $\mathbf{V}$ and let $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be disjoint sets of variables.

(a) Prove that there exists a set $\mathbf{Z} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$ if and only if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}}$, where $\mathbf{Z_0} = An(\mathbf{X} \cup \mathbf{Y})_{\mathcal{G}} \setminus (\mathbf{X} \cup \mathbf{Y})$.[1]

Hint: Keep in mind that this proof has two directions, one of which is quite simple. For the other, a possible strategy is to assume the existence of $\mathbf{Z}$ and proceed by contradiction. Along the proof it may be able to fix any particular path $\overline{p}$ between some $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ and consider the types of triplets that it may contain.

(b) Let $\mathbf{R} \subseteq \mathbf{V} \setminus (\mathbf{X} \cup \mathbf{Y})$ and $\mathbf{I} \subseteq \mathbf{R}$. Prove that there exists a set $\mathbf{Z}$, $\mathbf{I} \subseteq \mathbf{Z} \subseteq \mathbf{R}$ such that $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z})_{\mathcal{G}}$ if and only if $(\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z_0})_{\mathcal{G}}$, where $\mathbf{Z_0} = An(\mathbf{X} \cup \mathbf{Y} \cup \mathbf{I})_{\mathcal{G}} \cap \mathbf{R}$.

[1] The set $An(X)_{\mathcal{G}}$ is defined as $\{V \in \mathbf{V} \mid \exists$ a path (possibly of zero length) $V \to \cdots \to X$ in $\mathcal{G}\}$. Then $An(\mathbf{X})_{\mathcal{G}} = \bigcup_{X \in \mathbf{X}} An(X)_{\mathcal{G}}$. Notice that $An(X)_{\mathcal{G}}$ includes X.