CAUSAL THINKING (MATH-352)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Exam Style Questions

Date: xxxxx
Time: xxxxx

Name: ________________________

SCIPER: ________________________

INSTRUCTIONS TO CANDIDATES

- This exam will contribute 80% to your final grade. To obtain the maximum number of points you should be clear about your reasoning and present your arguments explicitly. You have **3 hours** to complete the exam.
- All that can be used for this exam is a pen. No notes, books, summaries, formula collections or calculators are allowed. All questions should be answered.
- The finest enumerated item in each question will be marked on a scale of $0-2$ points, indicating an incorrect, partially correct and completely correct answer respectively (half-points are not given). **The exam has xxx questions with a total of xxxxx points**.
- **Write the answer to every question in this booklet**, in the blank spaces after the question, or in the blank spaces at the end of the booklet if you run out of space. Scrap paper will be provided for rough work, but only answers written in the booklet will be marked.

Mark question 1 (TOT: xxxxx points):

Mark question 2 (TOT: xxxxx points):

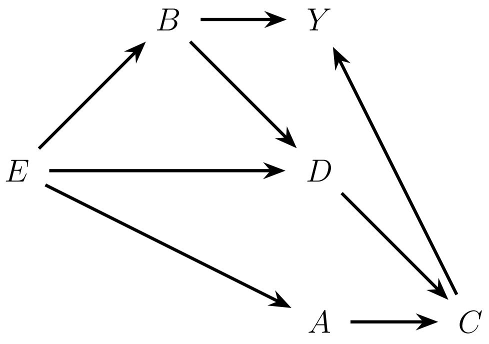
Mark question 3 (TOT: xxxxx points):

Mark question 4 (TOT: xxxxx points):

Mark question 5 (TOT: xxxxx points):

Mark question 6 (TOT: xxxxx points):

Question 1 (DAG warm-up)**.** Consider the DAG below.



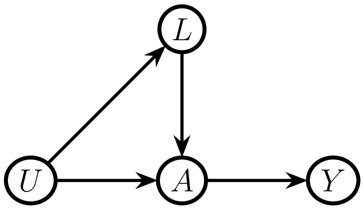
Use the rules of d-separation to decide whether the following independencies hold. Justify your answer:

(1) $D \perp\!\!\!\perp A | E$,
(2) $Y \perp\!\!\!\perp E | B, C$
(3) $Y \perp\!\!\!\perp A | C, E$

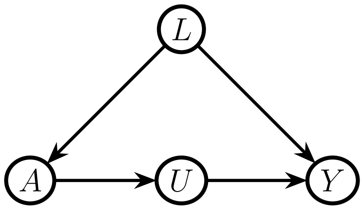
Suppose we are interested in the causal effect of A on Y and C. Are the following statements true or false? Justify your answer.

(4) $C^a \perp\!\!\!\perp A | D, Y^a$
(5) $Y^a \perp\!\!\!\perp A | E$

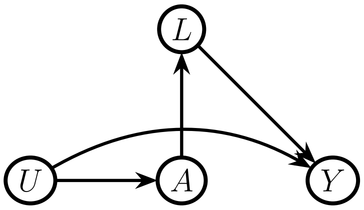
Question 2 (Graphical independencies)**.** Let L, U, A, Y denote baseline covariates, an unmeasured variable, treatment and outcome (temporally and topologically ordered as written, with L first and Y last). Statisticians 1, 2 and 3 propose the causal models $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$ respectively:



$\mathcal{G}_1$

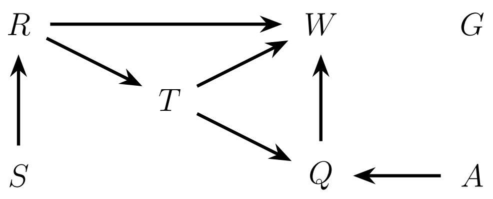


$\mathcal{G}_2$



$\mathcal{G}_3$

(1) Does $\mathcal{G}_1$ or $\mathcal{G}_2$ imply any independencies in the observed law $p(l, a, y)$? Can the observed law $p(l, a, y)$ be used to falsify $\mathcal{G}_1$ or $\mathcal{G}_2$?
(2) Determine the causal effect $E[Y^{a=1} - Y^{a=0}]$, assuming that consistency and positivity holds, and that
  (a) $\mathcal{G}_1$ is correct.
  (b) $\mathcal{G}_2$ is correct.
  (c) $\mathcal{G}_3$ is correct.
  In your answer for parts (b) (i)–(iii), express $E[Y^{a=1} - Y^{a=0}]$ as a function of the observed law $p(l, a, y)$ if this is possible, otherwise explain why it is not possible.

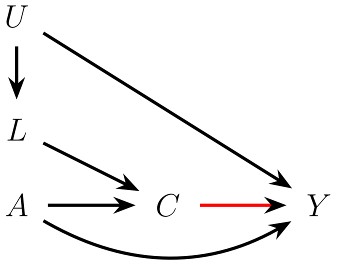


Question 3. In the DAG above, R represents area of residence, S represents socioeconomic status, T represents traffic noise, W represents subjective well-being, Q represents sleep quality, G represents sex and A represents age.

(1) Which statements are correct, supposing that all regression models are correctly specified and all relationships between variables are linear, without interactions?
  (a) In a regression of 'subjective well-being' on 'age' and 'sex', we expect both coefficients to be zero.
  (b) The coefficients in a regression of 'sleep quality' on 'traffic noise', 'subjective well-being' and 'socioeconomic status' have a causal interpretation.
  (c) The coefficients in a regression of 'subjective well-being' on 'traffic noise' and 'sleep quality' have a causal interpretation.
  Note: We say the coefficients have a "causal interpretation" if they can be interpreted as a (counterfactual) causal estimand.
  (d) If we regress 'sleep quality' on 'age' and obtain a coefficient of about zero, we know that the DAG is wrong.

(2) Suppose we are interested in the total causal effect of 'sleep quality' on 'subjective well-being'. For what confounders do we need to adjust? How could the desired effect be obtained?

Question 4 (SWIGs and identification)**.** Consider the DAG below. Unless directed otherwise, ignore the **red** arrow in the graph (pretend it is not there, until told otherwise). Furthermore, you may assume that all nodes are binary with a positive probability for each value, and that consistency holds.



Suppose you are consulting with a clinician who is interested in estimating the expected potential outcomes in their study population under treatment level $a = 1$ and comparing it to that under treatment level $a = 0$. Suppose the investigator has access to data on $\{L, A, Y\}$ but that C represents a variable indicating that patients were lost to follow-up and data on Y for patients with $C = 1$ is not recorded.

(1) Consider an intervention that sets treatment $A = a$. What are the backdoor paths for this intervention with respect to outcome Y?

(2) Is the backdoor criterion satisfied? By drawing the SWIG $\mathcal{G}(a)$, check whether exchangeability ($Y^a \perp\!\!\!\perp A$) holds. If so provide a functional of parameters of the distribution of $\{L, A, Y\}$ that identifies $\mathbb{E}[Y^a]$. If so, describe an estimation strategy for this functional that you would share with your collaborators, or name some issues that would arise.

(3) Draw the SWIG $\mathcal{G}(a, c = 0)$ corresponding to an intervention that sets A to a and C to 0.

(4) Do the independencies $Y^a \perp\!\!\!\perp A$ and $Y^a \perp\!\!\!\perp C^a \mid A, L$ hold? Justify your answer. If so provide a functional of parameters of the distribution of $\{L, A, Y\}$ that identifies $\mathbb{E}[Y^{a,c=0}]$ using the identification theorem for the g-formula or otherwise. If so, describe an estimation strategy for this functional that you would share with your collaborators.

(5) This parameter has a different interpretation than the one that your collaborators first described. Can you convince them that this parameter is still of interest, and if so, how?

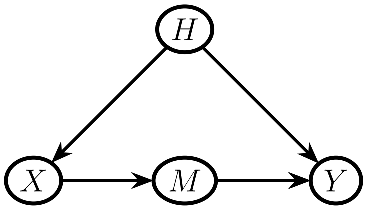
(6) Suppose now that the red arrow is present. Repeat the following steps for $\mathbb{E}[Y^{a,c=0}]$ using this modified graph:

(a) (c)

(b) (d)

(c) (e)

Question 5 (Identification in another graph)**.** Assume that variables X, M, Y satisfy the causal model $\mathcal{G}$ below, in which we take H to be an unmeasured variable. Furthermore, you may assume that all variables are discrete, and Y is binary with support $\{0, 1\}$.



(1) Investigator 1 suggests the following identification formula (g-formula) for $E[Y^x]$:

$$E[Y^x] = E[Y \mid X = x] \; .$$

Argue whether this identification formula holds or fails.

(2) Investigator 2 suggests another identification formula for a causal effect:

$$P(Y^x = 1) = \sum_m P(M = m \mid X = x) \sum_{x'} P(Y = 1 \mid X = x', M = m) P(X = x') \; .$$

Show whether the identification formula holds or fails. You may assume that interventions on M are well-defined. Is this a g-formula?

Hint: Use exclusion restriction and independencies in SWIGs $\mathcal{G}(x)$, $\mathcal{G}(x, m)$ and $\mathcal{G}(m)$. Furthermore, you may use the fact that $Y^x = Y^m$ when $M^x = m$, since X only affects Y through the mediator M.

(3) Argue that

$$E[Y^x] = E\left[\frac{\pi(M \mid X = x)}{\pi(M \mid X)} Y\right]$$

where we define π in the usual way as $\pi(\bullet \mid \circ) = P(M = \bullet \mid X = \circ)$.

(4) State the positivity condition which is required for the identification formula in (c) to be well-defined.

Question 6. *Challenge.* In this example, we will consider the effect of smoking (considered as binary variable A) on lung cancer (binary variable Y). It has been observed that these variables are correlated, and that the associational risk ratio comparing the risk in those who smoke against the risk in those who do not smoke, is $RR_{Y|A} = \frac{P(Y=1|A=1)}{P(Y=1|A=0)} = 10.73$.

Ronald Fisher argued that the observed correlation between A and Y may be due to some unknown genetic variant (binary variable U), which both causes smoking and causes cancer. In response, Jerome Cornfield argued that if U completely explains away the association between A and Y, such that $Y \perp\!\!\!\perp A \mid U$, it must also hold that $RR_{Y|U} = \frac{P(Y=1|U=1)}{P(Y=1|U=0)} \geq RR_{Y|A}$ and $RR_{U|A} = \frac{P(U=1|A=1)}{P(U=1|A=0)} \geq RR_{Y|A}$. Justify this result.

Hint 1: use laws of probability, e.g. start by rewriting $RR_{Y|A}$ as

$$\frac{\sum_u P_{Y|A,U}(1|1,u) \times P_{U|A}(u|1)}{\sum_u P_{Y|A,U}(1|0,u) \times P_{U|A}(u|0)}\ .$$

Hint 2: observe that if $RR_{Y|A} = 10.73$ then without loss of generality, we can assume that $P_{U|A}(1 \mid 1) \geq P_{U|A}(1 \mid 0)$ and $R_{U|Y} \geq 1$.

References