Exercises

1	2	3	4	5	6

Surname, First name

1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9
0	0	0	0	0	0	0	0

Causal Data Science (5294CADS6Y)
Practice exam

Week 1: Explaining the potential and limitations of causal inference, basics

1p **1a** Correlation is ____ causation

1p **1b** If X causes Y and Y causes X we say this causal relationship forms a ______

1p **1c** In cases similar to the Simpsons' paradox, does one need to know the causal graph in order to know which variables to adjust for when estimating a causal effect?

(a) Yes (b) No

1p **1d** In a randomized control trial, how are the treatment and control group chosen

(a) We choose randomly (e.g. with a coin flip)
(b) We choose to treat the patients that are more likely to get a good outcome

1p **1e** We know aspirin is an effective treatment for headaches, which means that taking it will improve the chances of a positive outcome. A patient had a headache, took an aspiring and now is feeling ok. What would have happened if they didn't take the aspirin?

(a) We cannot say what would have happened from this data alone
(b) They would have been cured
(c) They wouldn't have been cured

1p **1f** In a given distribution P, X is independent of Y. Which of the following also always holds in P?

(a) X is independent of Y given Z
(b) Y is independent of X
(c) X is independent of Z

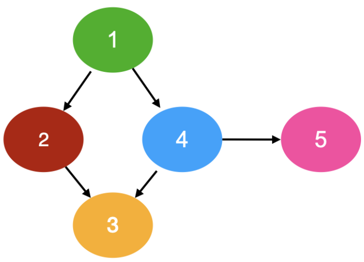
5p **1g** We have 3 random variables A, B, C, for which $A \perp\!\!\!\perp C \mid B$
Write down all 5 simplest factorizations that you can get from the chain rule and then applying the conditional independence simplification when it is possible.

Week 2: Graphical models, d-separation and causal graphs

1p **2a** Given a graph, a path between node i and node j is a sequence of distinct nodes (i, ..., j), such that each two consecutive nodes are ____________________

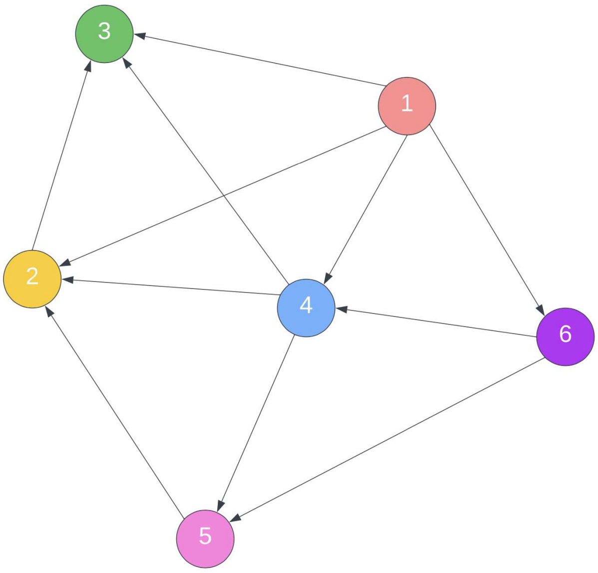
1p **2b** A collider on a path is an node with two ____________________ edges

1p **2c** In the following graph is the path (1,4,3) blocked when I condition on 4 and 5?



(a) Yes (b) No

10p **2d** Which are the d-separations that hold in this graph? List **at least 5** d-separations that hold and for each explain briefly why it is holds in this graph.



5p **2e** Consider the following linear Gaussian structural causal model:

$$\begin{cases} X \leftarrow \epsilon_X \\ Y \leftarrow -2X + 2\epsilon_Y \\ Z \leftarrow \frac{X}{2} + \epsilon_Z \\ \epsilon_X, \epsilon_Y, \epsilon_Z \sim N(0,1) \\ \epsilon_X \perp\!\!\!\perp \epsilon_Y, \epsilon_Y \perp\!\!\!\perp \epsilon_Z, \epsilon_X \perp\!\!\!\perp \epsilon_Z \end{cases}$$

Hint: keep in mind that for two random variables A and B

$$\mathbb{E}[aA + bB] = a\mathbb{E}[A] + b\mathbb{E}[B]$$
$$\mathrm{Var}[aA + bB] = a^2\mathrm{Var}[A] + b^2\mathrm{Var}[B] \qquad \text{if} \quad A \perp\!\!\!\perp B$$

Compute the following distributions, where the last question represents the distribution of Y after a soft intervention on X that changes its mechanism to $X \leftarrow 2 + \epsilon_X$

1. $P(Y) =$
2. $P(Y|do(Z = 1)) =$
3. $P(Y|do(X = 3)) =$
4. $P_{\text{soft-intervention}:X \leftarrow 2+\epsilon_X}(Y) =$

Additionally explain how you calculated the answer for question 4:

Week 3: Identification strategies for causal effect estimation

1p **3a** List at least four identification strategies that we have studied in class:

1p **3b** Any path between the treatment and the outcome, that starts with an edge pointing into the treatment, is a ____________________ path

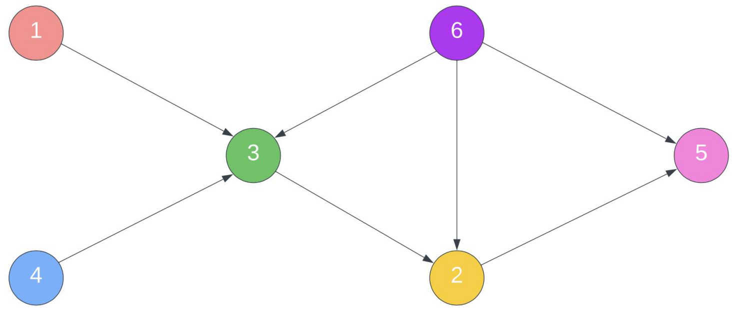
1p **3c** The frontdoor criterion combines the application of which other criteria: 1) backdoor for the effect of the treatment on the mediator and backdoor for the effect of the mediator on the outcome, 2) IV for the treatment and backdoor for the effect of the mediator on the outcome, 3) adiustment criterion for the effect of the treatment on the mediator and IV for the effect of the mediator on the outcome

- (a) Adjustment criterion for the effect of the treatment on the mediator and IV for the effect of the mediator on the outcome
- (b) IV for the effect of the treatment on the mediator and backdoor for the effect of the mediator on the outcome
- (c) Backdoor for the effect of the treatment on the mediator and backdoor for the effect of the mediator on the outcome

1p **3d** Which of the following assumptions is one of the requirements for a variable I to be considered a valid instrument?

- (a) I blocks all backdoor paths from the treatment to the outcome
- (b) I mediates all effects from the treatment to the outcome
- (c) I does not cause the treatment directly

5p **3e** Given the following graph apply the backdoor criterion for the effect of 3 on 2 and find all valid adjustment sets according to the backdoor criterion, explaining why they are valid.



5p **3f** **Difficult:** Discuss the applicability of the backdoor criterion, the frontdoor criterion and instrumental variables. Which are the cases in which one can apply only some of these criteria?

Week 4: Causal effect estimation - Matching, propensity scores, IPW

1p **4a** What is the formula for the average effect of treatment on the treated (ATT)?

1p **4b** Which of the following assumptions are typical in the potential outcome framework (**CHOOSE ALL THAT APPLY):**

- ☐ (Conditional) ignorability (no unmeasured confounding)
- ☐ Positivity (non-deterministic assignment)
- ☐ Interference (network effects)
- ☐ Stable Unit Treatment Value Assignment
- ☐ Consistency
- ☐ Linearity (linear SCM)

1p **4c** The propensity score is:

- (a) The inverse of the the probability of getting assigned the treatment T
- (b) The probability of getting assigned the treatment T given a certain value for the covariates X
- (c) The probability of a certain value for the covariates X given the treatment is assigned

5p **4d** Given the SCM below and the realization for a specific unit i, compute the unit-level counterfactuals for this unit for the following two cases, explaining every step:
(1) X=0 and
(2) Y=0

$$
\begin{aligned}
X &\leftarrow -1 + \epsilon_X & \epsilon_X &\sim \mathcal{N}(0,2) \\
Y &\leftarrow -X + \epsilon_Y & \epsilon_Y &\sim \mathcal{N}(0,1) \\
Z &\leftarrow X + Y + \epsilon_Z & \epsilon_Z &\sim \mathcal{N}(0,0.1) \\
V &\leftarrow 3Y + \epsilon_V & \epsilon_V &\sim \mathcal{N}(0,1)
\end{aligned}
$$

Realization for a specific unit i:

$$(x^i = -0.9, \quad y^i = 0.7, \quad z^i = -0.2, \quad v^i = 2.2)$$

5p **4e** **Difficult:** Discuss the differences between matching and propensity score matching, and their relative benefits or drawbacks.

Week 5: Causal Discovery

1p **5a** What type of structure does the PC algorithm output? ______________________

5p **5b** Apply SGS on this set of conditional independences and show the results of each phase, explaining why you remove or orient edges at each step.

$$1 \perp\!\!\!\perp 3$$

$$1 \perp\!\!\!\perp 5|4$$

$$2 \perp\!\!\!\perp 5|4$$

$$3 \perp\!\!\!\perp 5|4$$

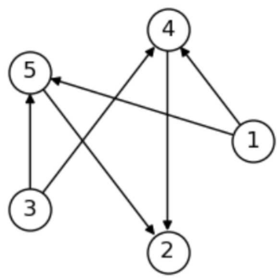
$$1 \perp\!\!\!\perp 4|2, 3$$

$$1 \perp\!\!\!\perp 5|2, 3$$

2p **5c** In score-based causal discovery, all the graphs with the same score should be part of the same Markov Equivalence Class.

(a) True (b) False

10p **5d** Starting from this CPDAG, which are the other CPDAGs in the phase 1 neighbours of GES of this CPDAG?
Explain how you computed the results step by step.



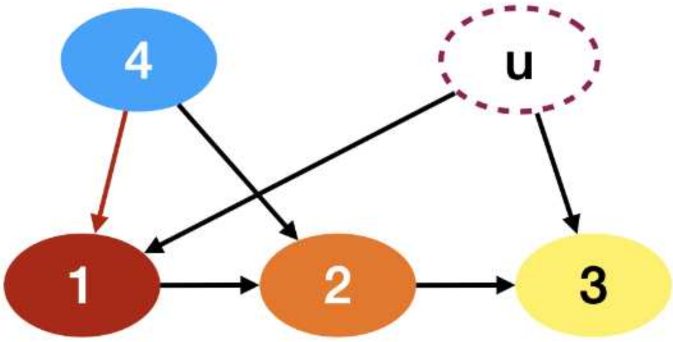
10p **5e** **Difficult:** Discuss the applicability and trade-offs between constraint-based, score-based causal discovery methods and restricted models like Additive noise models or LINGAM.

Week 6: do-calculus, transportability, advanced structure learning

1p **6a** Which method is complete in terms of identification? ____________________

2p **6b** To compute transportability we need to encode the changes in the different datasets in terms of __________ diagrams

3p **6c** $P(X_3|X_1)$In this graph and any distribution P that is Markov and faithful to it,$P(X_3|do(X_1))$ is :



- (a) P(X3|X1)
- (b) P(X3)
- (c) Neither P(X3) nor P(X3|X1)

10p **6d** **Difficult:** Discuss the relationship between causal discovery and identification strategies, including potential limitations and ideas on how to combine these two areas.