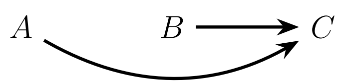
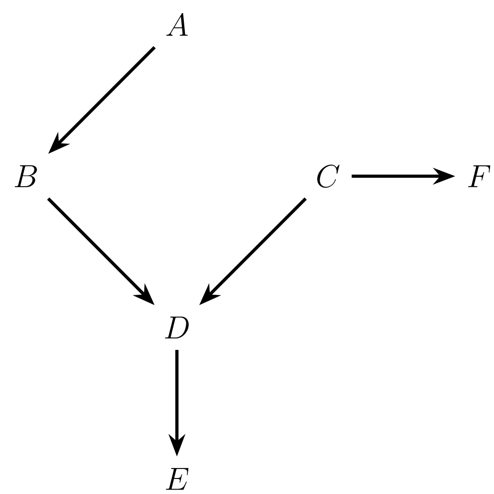
EXERCISES FOR RANDOMIZATION AND CAUSATION (MATH-336)

Exercise Sheet 3

Exercise 1 (DAGs and independencies)**.**



(A)



(B)

(a) Write down the Markov factorization for:
   (i) $p(a, b, c)$ in DAG (A),
   (ii) $p(a, b, c, d, e, f)$ in DAG (B).

(b) Now imagine that DAGs (A) and (B) were complete and were ordered alphabetically. In other words, A receives no edges; B receives an edge from A; C receives edges from A, B etc. Write down the Markov factorizations for:
   (i) $p(a, b, c)$ in DAG (A),
   (ii) $p(a, b, c, d, e, f)$ in DAG (B).

(c) By comparing your answers to (a) and (b) factor-by-factor, determine the independencies implied by each of the DAGs. These are called the defining (conditional) independencies of the DAG.

Exercise 2 (Faithfulness)**.** Suppose a law $\mathbb{P}$ is faithful to a DAG G. In the following you are given a complete list of independecies for the random variables involved. Find all the graphs G that satisfy the conditions

(a) $X \perp\!\!\!\perp Z$ for variables (X, Y, Z).

(b) $X \perp\!\!\!\perp Y | Z$ for variables (X, Y, Z).

(c) $X \perp\!\!\!\perp Y$,
   $X \perp\!\!\!\perp W | Z$,
   $X \perp\!\!\!\perp W | Z, Y$,
   $Y \perp\!\!\!\perp W | Z$,

$Y \perp\!\!\!\perp W | Z, X$,
for variables (X, Y, Z, W).

Exercise 3 (Collider paths)**.** The Graduate Record Examinations (GRE), a set of standardized tests, are commonly used to assess applicants for graduate programs in the United States. A study was conducted to investigate whether GRE test scores could be used to predict various performance outcomes among graduate students [1]. For the quantitative GRE (the mathematics part of the exam), analyses such as Fig. 1 were performed, and it was found that quantitative GRE score is not a very good predictor of graduate student performance. We will now use causal reasoning to investigate a possible reason for this finding.

Denote the quantitative GRE test score by G, performance outcome (for example time to first author publication count) by Y and admission decision to graduate school by D. Furthermore, denote a person's quantitative skills by A_1, and denote other factors of success (for example scientific creativity, prior engagement in area of research interest etc) by the variable A_2. The estimands studied in Fig. 1 are thus of the form

$$E[Y \mid G, D = 1] \ .$$

Suppose that a PhD student's performance is described by the following structural equations:[1]

$$\begin{aligned} Y &= f_Y(A_1, A_2, D, U_Y) \\ D &= f_D(G, A_2, U_D) \\ G &= f_G(A_1, U_G) \\ A_1 &= f_{A_1}(U_{A_1}) \\ A_2 &= f_{A_2}(U_{A_2}) \ . \end{aligned}$$

Assume that the error terms $U_Y, U_G, U_D, U_{A_1}, U_{A_2}$ are mutually independent, and thus we have defined a NPSEM-IE.

Answer the following:

(a) Draw the causal DAG corresponding to the above structural equation system.
(b) Determine whether the following independencies hold in the DAG you created in (a):
  (i) $G \perp\!\!\!\perp A_2$
  (ii) $G \perp\!\!\!\perp A_2 \mid D$
(c) Suppose we discretize G and Y into binary caterogies such that $G \in \{0, 1\}$ (0 = lower test score, 1 = higher test score) and $Y \in \{0, 1\}$ (0 = weaker performance, 1 = stronger performance). Use the answer to part (b) to give a story, using causal arguments, why

$$E[Y \mid G = 1, D = 1] - E[Y \mid G = 0, D = 1] \approx 0 \ .$$

Exercise 4 (Yellow fingers and lung cancer)**.** Consider the following structural equation systems for treatment A, outcome Y and covariates L, all assumed to be binary with marginally independent errors $U_A \perp\!\!\!\perp U_L \perp\!\!\!\perp U_Y$ (this defines an NPSEM-IE causal model):

[1] This is not an entirely realistic assumption. We will return to the problem of evaluating such an assumption formally in a later exercise.

(a) Suppose that

$$Y = f_Y(A, U_Y) \; ,$$
$$A = f_A(U_A) \; .$$

(i) Draw the causal graph for the above causal model.

(ii) Compute the observed law $P(A = a, Y = y)$ given that

$$f_Y(A, U_Y) = A \cdot I\left(U_Y \leq \frac{5}{8}\right) + (1 - A) \cdot I\left(U_Y \leq \frac{3}{8}\right) \; ,$$
$$f_A(U_A) = I\left(U_A \leq \frac{1}{2}\right) \; ,$$

with U_A, U_Y being i.i.d. uniform random variables on $[0, 1]$.

(iii) Using the observed law, compute $E[Y \mid A = 1] - E[Y \mid A = 0]$.

(iv) Does A cause Y in this model?

(b) Suppose that

$$Y = \widetilde{f}_Y(L, U_Y) \; ,$$
$$A = \widetilde{f}_A(L, U_A) \; ,$$
$$L = \widetilde{f}_L(U_L) \; .$$

(i) Draw the causal graph for the above causal model.

(ii) Compute the observed law $P(A = a, Y = y)$ given that

$$\widetilde{f}_Y(L, U_Y) = L \cdot I\left(U_Y \leq \frac{3}{4}\right) + (1 - L) \cdot I\left(U_Y \leq \frac{1}{4}\right) \; ,$$
$$\widetilde{f}_A(L, U_A) = L \cdot I\left(U_A \leq \frac{3}{4}\right) + (1 - L) \cdot I\left(U_A \leq \frac{1}{4}\right) \; ,$$
$$\widetilde{f}_L(U_L) = I\left(U_L \leq \frac{1}{2}\right) \; ,$$

with U_A, U_Y, U_L being i.i.d. uniform random variables on $[0, 1]$.

(iii) Using the observed law, compute $E[Y \mid A = 1] - E[Y \mid A = 0]$.

(iv) Does A cause Y in this model?

(v) Suppose that A had actually been randomized in the observed data (assume that the value of A was assigned by flipping an unbiased coin). How would the structural equations and causal graph in part (b) change?

(c) Deduce that the observed law of $P(A = a, Y = y)$ does not correspond to a single structural equation model. Thus, knowledge of $P(A = a, Y = y)$ is insufficient to determine whether A causes Y.[2]

(d) An investigator wants to conduct an experiment to test whether having yellow fingers causes lung cancer. To do so, she stops 10 individuals with yellow fingers and 10 individuals without yellow fingers on the street. Then, she asks them whether or not they have been diagnosed with lung cancer. She finds that $2/10$ individuals with yellow fingers have lung cancer, versus $1/10$ individuals without yellow fingers.

[2]This observation is frequently described by the aphorism 'correlation is not equal to causation'.

(i) Using your answers to part (a) and (b), suggest a causal story (draw a graph, define the nodes) which explains the relationship between smoking, yellow fingers and lung cancer (you can assume that smoking causes both yellow fingers and lung cancer).

(ii) Based on the graph, can we conclude from the observation in (c) that yellow fingers cause lung cancer? There is no need to perform any computations at this point: a full argument with counterfactuals will be the subject of a later question.

(iii) Would your answer to (c)-(ii) change if the numbers were different: 100 000 with with lung cancer amongst 1 000 000 persons without yellow fingers versus 200 000 with lung cancer amongst 1 000 000 persons with yellow fingers?

References

[1] Liane Moneta-Koehler, Abigail M. Brown, Kimberly A. Petrie, Brent J. Evans, and Roger Chalkley. The Limitations of the GRE in Predicting Success in Biomedical Graduate School. *PLOS ONE*, 12(1):e0166742, January 2017. Publisher: Public Library of Science.

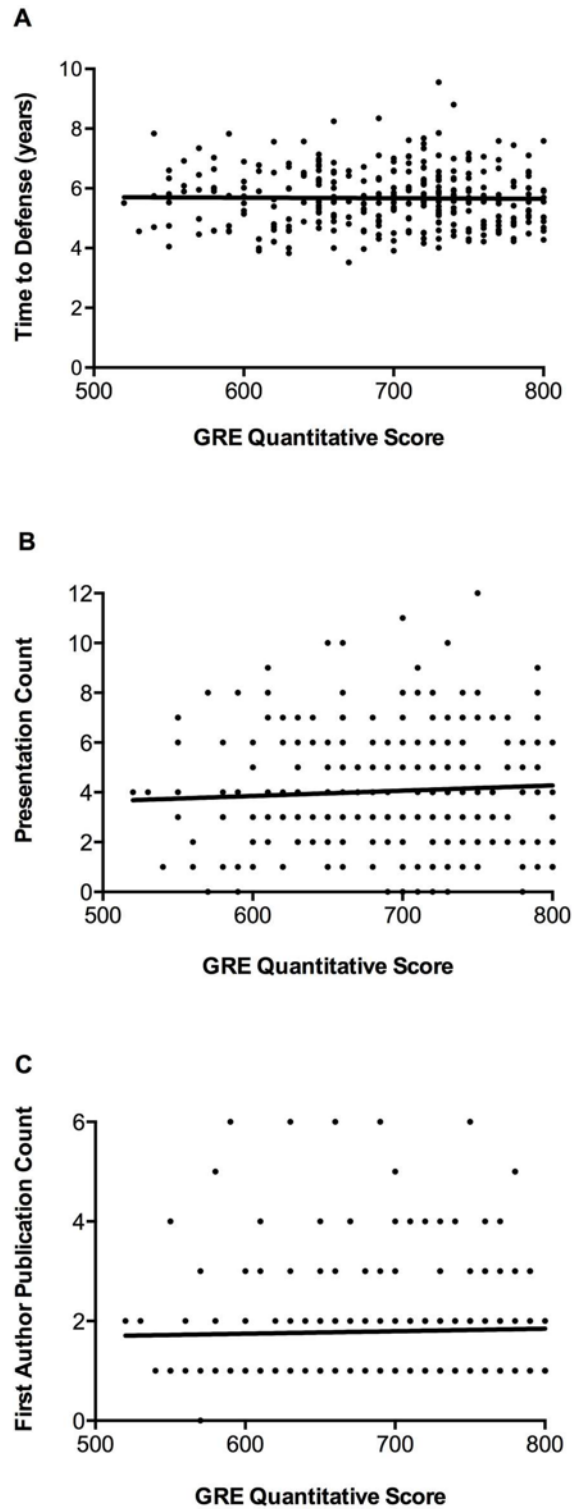


FIGURE 1. Association between performance outcomes in graduate school and and GRE test scores. Non-significant ($P \geq 0.05$) correlation coefficients were observed. Reproduced from [1].