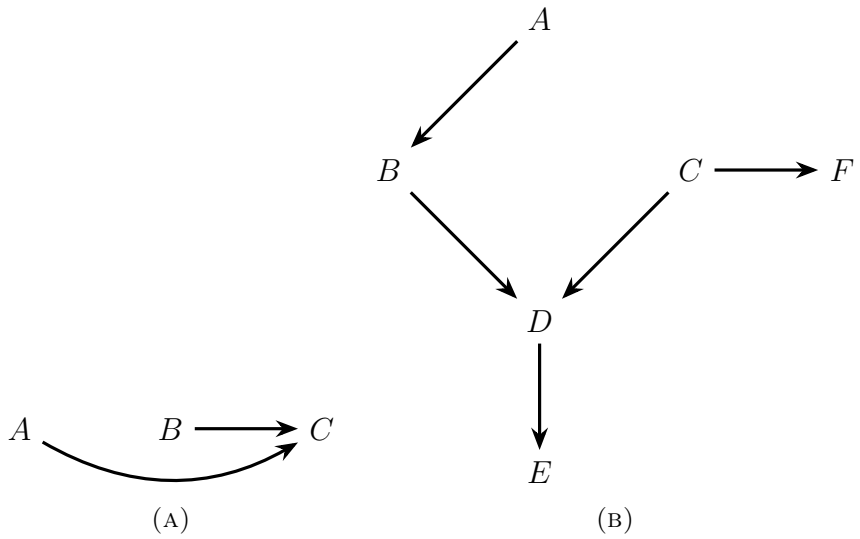
EXERCISES FOR RANDOMIZATION AND CAUSATION (MATH-336)

Exercise 1 (DAGs and independencies)**.** (from Robins' EPI 207 [1])



(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.

Solution:

(a) The Markov factorizations are given by
  (i) $p(a, b, c) = p(c \mid a, b)p(b)p(a)$,
  (ii) $p(a, b, c, d, e, f) = p(f \mid c)p(e \mid d)p(d \mid b, c)p(c)p(b \mid a)p(a)$.

(b) The Markov factorizations are given by
  (a) $p(a, b, c) = p(c \mid b, a)p(b \mid a)p(a)$,
  (b) $p(a, b, c, d, e, f) = p(f \mid a, b, c, d, e)p(e \mid a, b, c, d)p(d \mid a, b, c)p(c \mid a, b)p(b \mid a)p(a)$.

(c) The defining conditional independencies are given by
  - DAG A: $B \perp\!\!\!\perp A$ because $p(b \mid a) = p(b)$.
  - DAG B:
    - $F \perp\!\!\!\perp (A, B, D, E) \mid C$ because $p(f \mid a, b, c, d, e) = p(f \mid c)$.

– $E \perp\!\!\!\perp (A, B, C) \mid D$ because $p(e \mid a, b, c, d) = p(e \mid d)$.
– $D \perp\!\!\!\perp A \mid B, C$ because $p(d \mid a, b, c) = p(d \mid b, c)$.
– $C \perp\!\!\!\perp (A, B)$ because $p(c \mid a, b) = p(c)$.

Exercise 2 (Faithfulness)**.** Suppose a law $\mathbb{P}$ is faithful to a DAG G. In the following you are given a complete list of independencies 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).

Solution:

(a) The unique solution is the following graph:

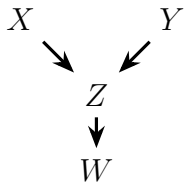
$$X \longrightarrow Y \longleftarrow Z$$

Note that X and Z are d-separated by the empty set. If we remove any of the edges in the graph, new indepenencies will be implied (which would violate the assumption that the list is complete).

(b) All the following graphs satisfy the conditions:

$$X \longrightarrow Z \longrightarrow Y$$
$$X \longleftarrow Z \longleftarrow Y$$
$$X \longleftarrow Z \longrightarrow Y$$

(c) The unique solution is the following graph:



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 [2]. For the quantitative GRE (the mathematics part of the exam), analyses such as Fig. 1 were performed, and the investigators concluded that the quantitative GRE score is a poor 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

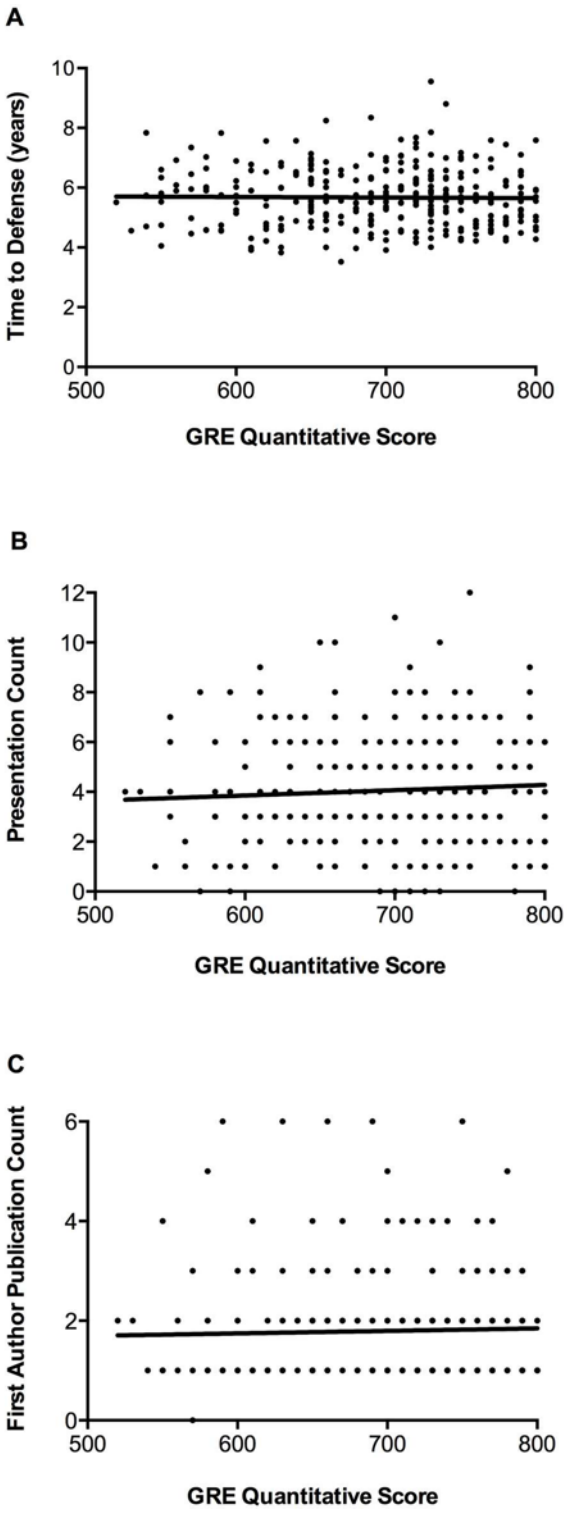


FIGURE 1. Association between performance outcomes in graduate school and and GRE test scores. Non-significant ($P \geq 0.05$) correlation coefficients were observed. Reproduced from [2].

(for example scientific creativity, or prior engagement in the area of research interest) by the variable A_2. The estimands studied in Fig. 1 are on 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.

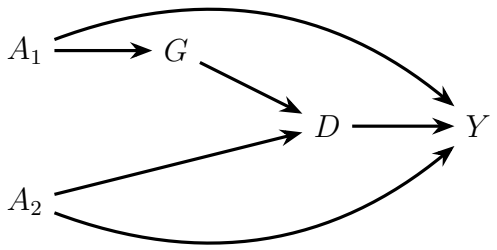
(b) Determine whether the following independencies hold in the DAG you created in (a):
   (i) $G \perp\!\!\!\perp A_2$
   (ii) $G \perp\!\!\!\perp A_2 \mid D$

(c) Suppose we discretize G and Y into binary categories 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

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

Solution:

(a) The causal DAG is given below (as usual, we omit the error terms to avoid clutter):



(b) (i) $G \perp\!\!\!\perp A_2$ because there is no open path from G to A_2 (the path $G \leftarrow A_1 \rightarrow Y \leftarrow A_2$ is blocked by the collider Y and $G \rightarrow D \leftarrow A_2$ is blocked by the collider D).
   (ii) $G \not\perp\!\!\!\perp A_2 \mid D$ because conditioning on D opens the collider path $G \rightarrow D \leftarrow A_2$.

(c) The fact that $A_1 \not\perp\!\!\!\perp A_2 \mid D$ can be written as

$$E[A_2 \mid G = 1, D = 1] \neq E[A_2 \mid G = 0, D = 1] \ .$$

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

This tells us that the strata $\{G = 1, D = 1\}$ and $\{G = 0, D = 1\}$ contain individuals with different A_2. In particular, it is conceivable that individuals admitted to graduate school with lower GRE scores, i.e. $\{G = 0, D = 1\}$, excel in other areas which have been recognized in the admission process. Conversely, admitted individuals with high GRE scores, i.e. $\{G = 1, D = 1\}$, may be weaker with regards to the factors A_2 than those admitted with low GRE scores. In words, low G implies high A_2 and vice versa when conditioning on $D = 1$. However, we saw that G and A_2 were marginally independent in (b)-(i).

Thus we can conclude that admitted candidates are either strong in A_1 or in A_2 (or both). Assuming that abilities A_1 and A_2 are of similar importance for performance outcomes Y, we then have that

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

Therefore, the predictive power of GRE scores on performance outcomes may be obscured by the presence of other independent factors for success in graduate school. A more informative way of assessing the predictive power of GRE scores would be to conduct an experiment where all applicants are admitted to graduate school, and then evaluate $E[Y \mid G = 1] - E[Y \mid G = 0]$ in this population, but this is clearly impractical.

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 mutually error terms, $U_A \perp\!\!\!\perp U_L \perp\!\!\!\perp U_Y$ (this defines an NPSEM-IE causal model):

(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.

(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?

Solution:

(a) (i) The causal graph is as follows:

$$A \longrightarrow Y$$

(ii) The observed law is given by

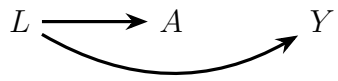
$$P(A = 0, Y = 0) = \frac{5}{16},$$
$$P(A = 0, Y = 1) = \frac{3}{16},$$
$$P(A = 1, Y = 0) = \frac{3}{16},$$
$$P(A = 1, Y = 1) = \frac{5}{16}.$$

(iii) Using (a)-(ii), we get that $E[Y \mid A = 1] - E[Y \mid A = 0] = \frac{1}{4}$.

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

(iv) In this causal model, A causes Y because there is a causal path (a directed path, i.e. a path where all edges follow the same direction) from A to Y in the graph.

(b) (i) The causal graph is as follows:



(ii) The law is the same as the one found in (a)-(ii).

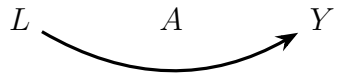
(iii) The contrast of expectations is the same as in (a)-(iii).

(iv) In this model, A does not cause Y because there is no causal path from A to Y.

(v) If A was randomized, it would no longer be affected by the value of L. Thus,

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

and the corresponding causal graph would be:



(c) Two different causal models can give the same observed distribution of A and Y, as seen in parts (a) and (b). In the former, A causes Y whereas in the latter, A does not cause Y. Therefore, it is necessary to specify a causal model (this can be done equivalently through structural equations and graphs) for the relationship between A and Y in order to decide whether A causes Y, i.e. whether an intervention on A would lead to a change in Y.

(d) (i) Smoking is a common cause of yellow fingers and lung cancer. This scenario is described by the structural equations and DAG in part (b), taking L to be smoking, A to be yellow fingers and Y to be lung cancer. Amongst smokers, there is an increased proportion of individuals with yellow fingers compared to non-smokers. Likewise, there is an increased proportion of individuals with lung cancer amongst smokers as compared to non-smokers. This creates an association between yellow fingers and lung cancer, even though yellow fingers does not in itself cause lung cancer.

(ii) The observed association between yellow fingers and lung cancer is likely due to confounding by smoking. Taking the observed contrast $E[Y \mid A = 1] - E[A \mid A = 0]$ as a measure of the causal effect of yellow fingers on lung cancer would therefore lead to a causal error. Moreover, we expect to see statistical error from random variability in the sample, because it is of small size ($n = 10$).

(iii) The new numbers would reduce the statistical error, but not the causal error.

References

[1] J. M. Robins. EPI 207 (Harvard T.H. Chan School of Public Health).

[2] 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.