Exercises MLSS 2019: *Causality*

Joris Mooij

August 27, 2019

1 In-class exercises

1.1 Simpson's Paradox

You are investigating the effectiveness of a drug against a deadly disease. You are given access to data collected by health insurance companies about their customers. You divide the diseased customers into two groups: those that took the drug, and those that didn't take the drug. Some of the customers recovered, others unfortunately didn't recover. The reasons why some patients were treated and others were not, are unknown to you. You find the following numbers:

	Recovery	No recovery	Total	Recovery rate
Drug	20	20	40	...%
No drug	16	24	40	...%
Total	36	44	80	

a) *Calculate the recovery rates (in %) for both groups ("drug" and "no drug").*

b) *If you were diseased, would you take the drug, or not?*

Upon closer inspection of the data, you notice something peculiar when you group patients according to gender:

Males	Recovery	No recovery	Total	Recovery rate
Drug	18	12	30	...%
No drug	7	3	10	...%
Total	25	15	40	

Females	Recovery	No recovery	Total	Recovery rate
Drug	2	8	10	...%
No drug	9	21	30	...%
Total	11	29	40	

c) *Calculate the recovery rates (in %) for both groups ("drug" and "no drug"), for each subpopulation (males and females) separately.*

d) *In light of these numbers, would you take the drug if you were diseased, or not?*

e) *What would be your advice to a diseased patient with unknown gender?*

This phenomenon is known as "Simpson's paradox". A lot has been written about this paradox, but it dissolves once you recognize that you should not make the mistake of interpreting correlations as causations, as we'll see later today.

1.2 Paths, colliders, d-blocked paths and d-separation

Definition 1 (Paths, Ancestors) *Let $\mathcal{G}$ be a directed mixed graph.*

- *A **path** q in $\mathcal{G}$ is a sequence of adjacent edges in $\mathcal{G}$ in which no node occurs more than once.*
- *A path consisting of directed edges $X_{i_1} \to X_{i_2} \to X_{i_3} \to \cdots \to X_{i_k}$ that all point in the same direction is called a **directed** path.*
- *If there is a directed path from X to Y (or if $X = Y$), X is called a **ancestor** of Y.*
- *The ancestors of Y are denoted $\mathrm{an}_{\mathcal{G}}(Y)$, and include Y.*

Definition 2 (Colliders, Blocked Paths, d-separation) *Let $\mathcal{G}$ be a directed mixed graph, and q a path on $\mathcal{G}$.*

- *A **collider** on q is a (non-endpoint) node X on q with precisely two arrowheads pointing towards X on the adjacent edges:*

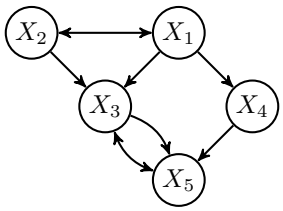
$$\to X \leftarrow, \quad \to X \leftrightarrow, \quad \leftrightarrow X \leftarrow, \quad \leftrightarrow X \leftrightarrow$$

- *A **non-collider** on q is any node on the path which is not a collider.*

*A set of nodes $\boldsymbol{S}$ in $\mathcal{G}$ is said to **d-block** q if q contains a non-collider which is in $\boldsymbol{S}$, or a collider which is* not *an ancestor of $\boldsymbol{S}$.*

*For three sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ of nodes in $\mathcal{G}$, we say that $\boldsymbol{X}$ **and** $\boldsymbol{Y}$ **are** d**-separated by** $\boldsymbol{Z}$ iff all paths between a node in $\boldsymbol{X}$ and a node in $\boldsymbol{Y}$ are d-blocked by $\boldsymbol{Z}$, and write $\boldsymbol{X} \perp_{\mathcal{G}} \boldsymbol{Y} \mid \boldsymbol{Z}$.*

Consider the following directed mixed graph $\mathcal{G}$:



a) *Is $X_3 \to X_5 \leftrightarrow X_3$ a path? Is it a directed path?*

b) *Is $X_3 \leftrightarrow X_5$ a path? Is it a directed path?*

c) *Is $X_5 \leftarrow X_3 \leftarrow X_1$ a path? Is it a directed path?*

d) *What are the ancestors of X_4?*

Consider the path $X_2 \leftrightarrow X_1 \to X_3 \leftrightarrow X_5 \leftarrow X_4$ on $\mathcal{G}$.

e) *Which nodes on the path are colliders?*

f) *Which nodes on the path are non-colliders?*

g) *Does $\{X_3\}$ d-block this path? Does $\{X_5\}$ d-block this path? Does $\{X_3, X_5\}$ d-block this path?*

h) *Does X_1 d-separate X_2 from X_4?*

i) *Is $X_1 \perp X_5 \,|\, \{X_3, X_4\}$?*