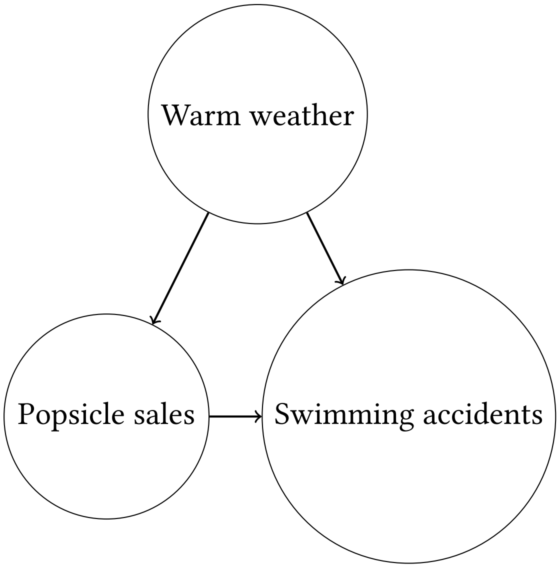
Data 102 (Fall 2022) **Names:**
Discussion Worksheet 8
10/14/2022

1. Brainstorm a possible confounder or collider in the following scenarios. Draw a causal graph for each scenario and label the treatment variable, the outcome variable and the confounder/collider.

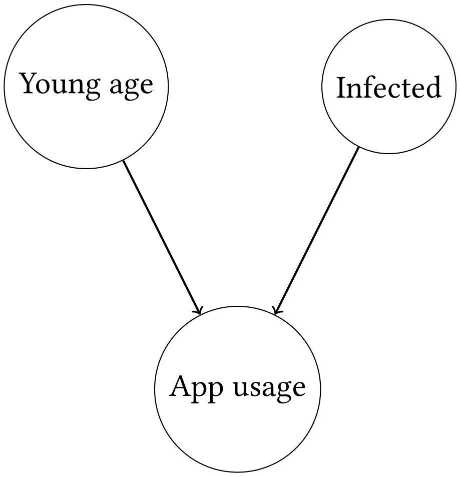
   (a) The Martian from lecture visits Earth again and concludes that popsicle sales cause swimming accidents.

   Solution: Warm weather is a confounder for popsicle sales and swimming accidents Y. Warm weather both increases popsicle sales and increases swimming activity, which increases swimming accidents. Not conditioning on warm weather generates a deceptive causal relationship between popsicle sales and swimming accidents.



   (b) There's a particular virus going around, and people can report their positive infection via an app. You collect data from the app and conclude that the virus has a higher chance of affecting 20-30 year olds than it does 60-70 year olds.

   Solution: App usage is a collider. Young people are more likely to use an app, hence the dependency between those two nodes, and if you are infected you can report on the app, hence the dependency between those two nodes. Conditional on using the app though, younger people appear to be more likely to be infected, whereas no such assumption is present for this scenario in reality.



2. A doctor performs a study on the relative effectiveness of two treatments, A and B, for kidney stones. She collects data on past patients that have been given treatments A and B to determine the relative success of the treatments. The results of the study are:

	Treatment A	Treatment B
# cases given treatment	100	100
# cases after treatment	25	20

   (a) What is the success rate of treatment A vs. treatment B?

   Solution: A: 0.75, B: 0.8

   (b) She investigates further to see how treatments A and B performed for large vs. small kidney stones and obtains the following data:

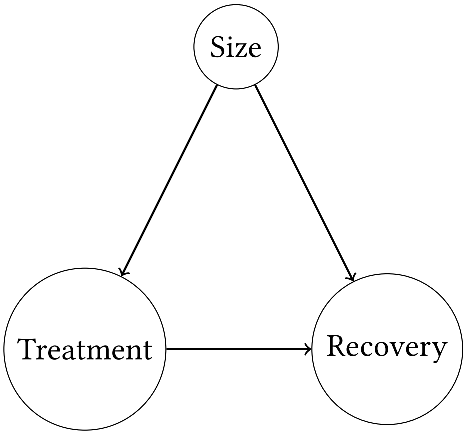
	Treatment A	Treatment B
# cases with small kidney stones given treatment	20	90
# cases with small kidney stones after treatment	2	15
# cases with large kidney stones given treatment	80	10
# cases with large kidney stones after treatment	23	5

   What are the success rates of treatments A and B for small and large kidney stones respectively?

   Solution: A. Small: $18/20 = 0.9$; Large: $57/80 = 0.7125$. B. Small: $75/90 = 0.83$; Large: $5/10 = 0.5$.

(c) How can you account for the fact that in aggregate B is a better treatment but for both small and large cases individually A is better? Is this an instance of Simpson's paradox or Berkson's paradox? What is the confounder or collider?

Solution: Possible causal story: treatment A is more severe (e.g. surgery) and therefore more effective on hard cases (like large stones), so perhaps more cases with large stones were assigned treatment A. Treatment B is more mild (e.g. medication). The confounder is size of the kidney stone, since not conditioning on the size of the kidney stones creates a deceptive causal relationship between treatment and recovery – i.e. makes it look like treatment B is more effective whereas in reality treatment A is more effective regardless of the size of the kidney stone. This is an example of Simpson's paradox – the paradox that arises when accounting vs. not-accounting for the confounder.

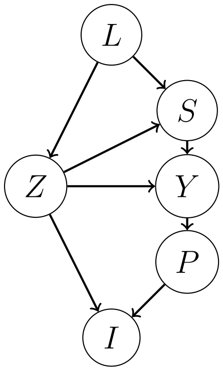


3. Consider the following variables:

- **L**: Location of garden
- **S**: Soil Quality
- **Z**: Rainfall (High or Low)
- **Y**: Number of flowers grown
- **P**: Total amount of Pollen on flowers
- **I**: Total number of Insects on flowers

(a) For the variables defined in the problem, draw the causal DAG which best captures their causal relationships.

Solution: We have the following causal DAG: L causes Z and S because climate and location determine soil quality and rainfall levels. S and Z have direct effects on the number of flowers that grow in a garden. Z impacts S and Y. Lastly Y causes P since more plants means more pollen, Z causes I because rain causes more insects to come out, and P causes I since more pollen means more insects.



(b) As we'll discuss in lecture, to measure causal effects we usually want to identify and condition on (adjust for) all confounding variables, while avoiding conditioning on colliders.

The *backdoor criterion* gives us a way to determine which variables are confounders. In particular, we simply need to "block" all the confounding pathways in the graphical model between two nodes.

In a causal graph, we define a *path* between two nodes X and Y as a sequence of nodes beginning with X and ending with Y, where each node is connected to the next with an edge (pointed in either direction). Given an ordered pair of variables (X, Y) – ordered pair here just means that the upcoming conditions focus on X, not Y – a set of variables S satisfies the *backdoor criterion* relative to (X, Y) if

- No node in S is a descendant of X (to prevent conditioning on colliders)
- S blocks every path between X and Y that contains an arrow into X.

Identify all sets of variables in the causal graph from part a) that satisfy the backdoor criterion relative to (Y, P).

Solution: You can find these sets by tracing all paths from Y to P, ignoring the directions of arrows. The first condition is that, if any node you encounter is a descendant of Y, it cannot be in the sets. The second condition is that each set must contain *at least one* node that blocks the path from Y to P. So, Z must be included in all sets, no set can include I, and L and S can be included or not. The sets are

$$\{Z\}, \{Z, S\}, \{Z, L\}, \{Z, S, L\}.$$

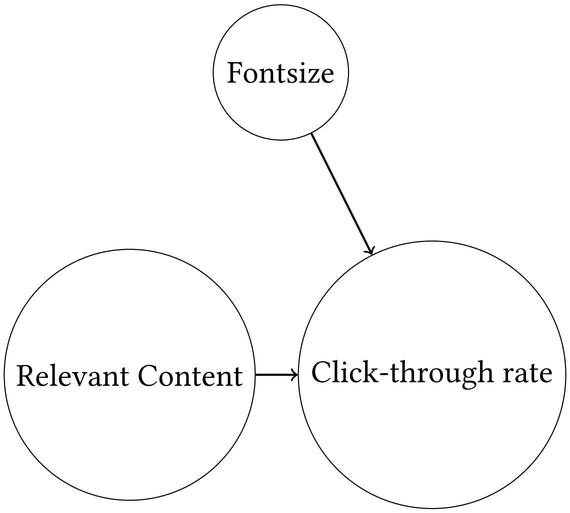
4. A search engine employs two data science teams to help display their content:

   - Team 1 determines where on the page to place content.
   - Team 2 determines what font size to use for each type of content.

The goal of team 1 is to place the content that is most relevant to a user highest on the page. The goal of team 2 is to give the content that generates the most ad revenue the largest font size.

Team 1 measures relevance based on how likely a user is to click on that item rather than other items (called the *click-through rate*). They discover that the font size of the content is a very predictive feature for click-through rate, and decide to add it to their model. Explain why this is a bad idea; draw and reference a causal graph as part of your explanation.

Solution: Fontsize is a confounder for click-through rate , so adding fontsize to the model will obscure the true effect that content placement has on click-through rate. People are more likely to click on more relevant content but also on content with larger font.



Feedback Form

On a scale of 1-5, where 1 = much too slow and 5 = much too fast, how was the pace of the discussion section?

1 2 3 4 5

Which problem(s) did you find most useful?

Which were least useful?

Any other feedback?