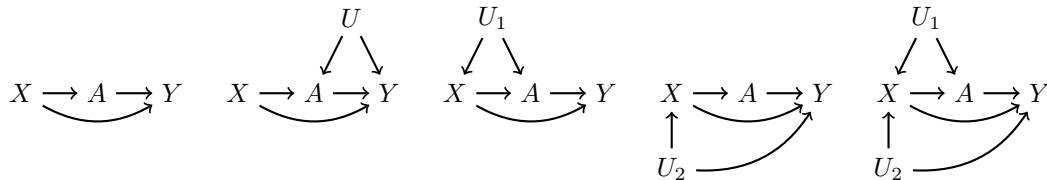


1 (32 points) Material covered Tuesday

Part 1 is about the **Directed Acyclic Graphs (DAGs)**.

For 1.1–1.5, answer True or False: X is a sufficient adjustment set to identify the causal effect of A on Y . Explain in one sentence. If False, state the backdoor path that is unblocked conditional on X .

- 1.1. (4 points) 1.2. (4 points) 1.3. (4 points) 1.4. (4 points) 1.5. (4 points)



For each statement below, answer True or False. Explain in one sentence.

- 1.6. (4 points) Conditioning on X blocks this path: $A \leftarrow B \leftarrow X \rightarrow C \rightarrow Y$
 1.7. (4 points) Conditioning on X blocks this path: $A \leftarrow B \rightarrow X \leftarrow C \rightarrow Y$

For the scenario below, draw a DAG with a counterexample. Explain to the researcher why this algorithm could produce misleading results.

- 1.8. (4 points) A researcher comes to you with a new machine learning method. It uses LASSO to search for variables that are predictive of both the treatment A and the outcome Y , and it includes in the model the union of those sets.

2 (18 points) Material covered Thursday

Part 2 is all about the **population inference**.

A researcher uses an opt-in online web survey to draw inference about support for President Biden. They ask respondents: “Do you approve of President Biden’s performance in office?” with the answer choices Yes/No. The researcher also gathers data on several demographic characteristics: race, whether the respondent completed college, and annual family income. They write:

The distribution of race, college, and income in my sample matches the distribution I estimate in the American Community Survey, a national probability sample collected by the Census Bureau. Therefore, my sample-based evidence about support for President Biden generalizes to the population.

This question is about formalizing a set of conditions under which the researcher is right and wrong. Assume throughout that the Census Bureau estimates are correct.

- 2.1. (5 points) Draw a DAG under which the researcher’s claim is valid. Use S as a random variable indicating inclusion in the sample.
 2.2. (3 points) In a sentence or two, explain your DAG from 2.1 to the researcher.
 2.3. (5 points) Draw a DAG showing a counterexample under which the researcher’s claim is invalid. Use S as a random variable indicating inclusion in the sample.
 2.4. (3 points) In a sentence or two, explain your DAG from 2.3 to the researcher.
 2.5. (2 points) One researcher does the above procedure with a sample of $n = 100$. Another researcher does the above procedure with a sample of $n = 1,000$. Does the size of the sample affect the validity of population inference?