CS 189 Introduction to Machine Learning
Fall 2018 **DIS11**

1 Simpson's Paradox

(For your convenience, we have reprinted the 2nd problem from last discussion on this worksheet.)

In 1973, overall admission rates to UC Berkeley graduate school displayed a significant gender imbalance (Figure 1), with male applicants being accepted more often than female applicants.

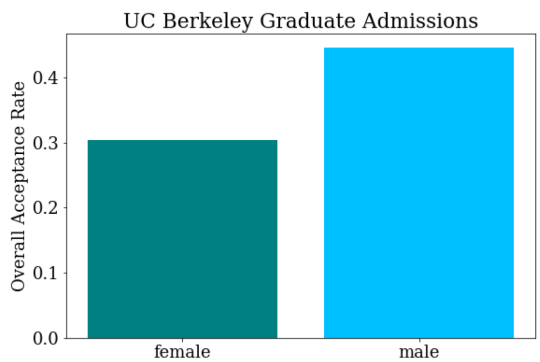

Figure 1: UC Berkeley Graduate Admissions by Gender

(a) Let Y be a random variable that denotes the admission decision (e.g. $Y = 1$ is the event of acceptance to graduate school). Let G be a random variable that denotes gender, which takes values in $\{\text{male}, \text{female}\}$. Use this notation to write the observation about overall acceptance rates as an inequality of probabilities.

(b) To investigate this problem, we look at the admissions practices of individual departments (Figure 2). Now it seems that the gender imbalance disappears or goes in the other direction!

Let D be a random variable that denotes the department, which takes values in $\{0, 1, 2, 3, 4, 5\}$. Use this notation to write the observation about acceptance rates by department as inequalities of probabilities.

(c) Write $P(Y = 1 \mid G = \text{female})$ in terms of $P(Y = 1 \mid G = \text{female}, D = i)$ for $i = 0, \ldots, 5$. Also write the expression for $P(Y = 1 \mid G = \text{male})$. Now, using the information in Figure 3, can you explain why the university-wide gender imbalance seems at odds with the pattern in individual departments?

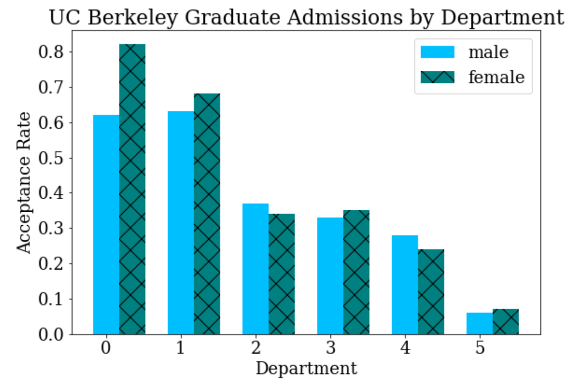

Figure 2: UC Berkeley Graduate Admissions by Department

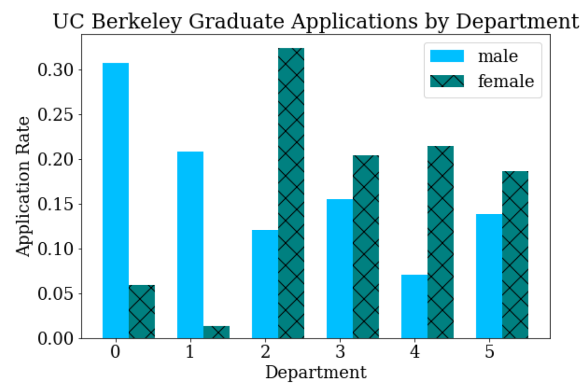

Figure 3: UC Berkeley Graduate Applications by Gender and Department

(d) This is an example of *Simpson's paradox*, which illustrates that drawing conclusions based on observational statistics may lead to incorrect conclusions. What does our statistical analysis suggest about the problem of gender imbalance overall?

2 Causal DAGs and Confounding

In Problem 2 of the previous discussion worksheet, we examined a 1973 study of the gender imbalance in UC Berkeley graduate school admission rates. The study found that the overall acceptance rate was lower for female applicants than for male applicants, that is,

$$P(Y = 1 \mid G = \text{male}) > P(Y = 1 \mid G = \text{female}) \tag{1}$$

The researchers weren't satisfied with this conclusion. They really wanted to answer the following *causal* question, "Did reporting 'female' on applications *cause* applicants to have lower overall acceptance rate?" This is a different question from "Was the overall acceptance rate different for female and male applicants?" (the answer to that is "yes". That's exactly what equation (1) tells us.)

In this problem, we will look at Simpson's paradox through the lens of causal inference. First, we begin with a review of causal directed acyclic graphs (DAGs).

Recall that a *directed graph* G consists of a set of vertices V and an edge set E of *ordered* pairs of variables. A directed acyclic graph (DAG) is a directed graph with no cycles.

We can represent joint probability distributions with DAGs. Let G be a DAG with vertices $X_1, ..., X_k$. If P is a (joint) distribution for $X_1, ..., X_k$ with (joint) probability mass function p, we say that G represents P if

$$p(x_1, \cdots, x_k) = \prod_{i=1}^{k} P(X_i = x_i | \text{pa}(X_i)), \tag{2}$$

where $\text{pa}(X_i)$ denotes the parent nodes of X_i. (Recall that in a DAG, node Z is a parent of node X iff there is a directed edge going out of Z into X.)

In other words, we can read off a *factorization* of the joint probability in terms of conditional probabilities, just based on how the nodes are connected in G.

(a) Consider the following DAG (Figure 4), G, which represents a joint distribution over X, Y, Z, denoted as $P_{X,Y,Z}$. Write down the factorization of $P_{X,Y,Z}$ that is represented by G.

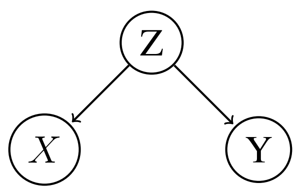

Figure 4: G, a DAG

(b) We call a DAG a *causal* DAG, if for any two nodes X, Y, there is an edge from X to Y iff X has a *direct causal effect*[1] on Y.

Consider a very simple 2-node model as follows.

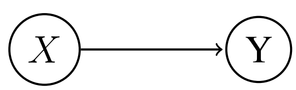

Figure 5: X has a direct causal effect on Y

One possible structural equation that corresponds to this model could be:

$$Y = aX + \epsilon, \tag{3}$$

where $a \in \mathbb{R}$ is the direct causal effect size and ϵ is a standard Gaussian random variable.

Suppose we could draw i.i.d. samples of X, Y from this model. How would you estimate the direct causal effect of X on Y?

[1] The formal definition of direct causal effect requires the introduction of the *do* operator and can get quite involved. For now we will operate on the level of intuition. Interested students may refer to Judea Pearl's book, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2nd edition, 2009.

(c) Now we add in a third node Z. In causal inference, Z is called a *confounder* because it has a direct causal effect on both the (potential) cause X and the outcome Y.

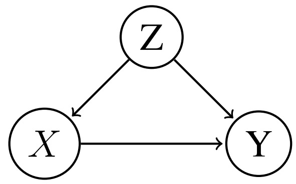

Figure 6: Confounder

One possible structural equation that corresponds to this model could be:

$$Y = aX + bZ + \epsilon_1, \tag{4}$$
$$X = cZ + \epsilon_2, \tag{5}$$

where ϵ_1, ϵ_2 are standard Gaussian random variables.

Suppose we could draw i.i.d. samples of (X, Y) from this model. Can you still use the 'naive' method you proposed in the previous part to estimate the direct causal effect of X on Y?

Suppose $a > 0$ (i.e. direct causal effect of X on Y is positive), give conditions on b, c under which the 'naive' method would suggest $a < 0$ instead.

(d) Suppose we could draw i.i.d. samples of (X, Y, Z) from this model. Propose a method to estimate the direct causal effect of X on Y.

(e) Recall the Simpson's paradox that you encountered in last week's discussion. Relate what you showed in the previous 2 parts to the Simpson's paradox.

