Problem Set 1

STAT II (Spring 2025)

Disclaimer: *Please read the guidelines below carefully. Good luck!*

Guidelines

- Please upload your answers to Brightspace by **Friday, February 7th, at 11:59 PM**.
- Direct all questions about the problem set to Brightspace under Contents > Discussions > Class Feed > Problem Set 1.
- For precise and expedited responses, we will address questions about this problem set on Brightspace until 6:00 PM on February 7th.
- Collaboration is allowed, but I encourage you to attempt the problems on your own before seeking help from others. Regardless of collaboration, you must individually write up and submit your answers.
- Late submissions will not be accepted unless prior approval is obtained from the instructor at least one day before the due date.
- **The total points for the problem set are 100**. Individual bonus questions may appear. If your total points exceed 100, the excess points will carry over to subsequent problem sets (e.g., if you earn 120 points, 20 points will be added to future problem sets).
- **Grading:** Show every step of your derivations. We grade the steps of derivations as well as the final answers. If you are unable to solve the problem completely, partial credit can be given for your derivations. Conversely, a correct final answer without complete derivations may not receive full credit.
- **Stylistic Requirements:** Adhere to the guidelines for the format of your submitted answers. Ensure you follow these rules:

    1. Submit your answers as a PDF file compiled from LaTeX, R Markdown, or (ideally) a Quarto Markdown document.
    2. If using raw LaTeX(.tex) for your answers, submit an accompanying .R file for any computational tasks, with referenced line numbers corresponding to each specific task.
    3. If using R Markdown (.Rmd) or Quarto Markdown (.qmd), include your code as code chunks in the source file. Additionally, submit the source .Rmd or .qmd file along with the compiled PDF to allow us to run your code easily.
    4. To ensure reproducibility of your simulation results, use `set.seed()` at the beginning of your document.

Problem 1 (10 points)

Suppose that X, Y, and Z are random variables. Prove the following:

$$\text{cov}(X + Y, Z) = \text{cov}(X, Z) + \text{cov}(Y, Z)$$

Problem 2 (10 points)

Suppose that you flip two fair coins. Define the random variables A, B, and C as follows:

$$A = \begin{cases} 1 & \text{if the first coin is heads} \\ 0 & \text{otherwise} \end{cases}$$

$$B = \begin{cases} 1 & \text{if the second coin is heads} \\ 0 & \text{otherwise} \end{cases}$$

$$C = \begin{cases} 1 & \text{if one coin is heads and the other is tails} \\ 0 & \text{otherwise} \end{cases}$$

Note that $A \perp\!\!\!\perp B$. Is it true that $A \perp\!\!\!\perp B \mid C$? Make sure to use the definition of conditional independence to offer a formal justification for your answer.

Problem 3 (30 points)

This problem aims to help you get familiar with the potential outcomes notation. Throughout the problem, you can assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

(a) Explain in words what the notation $Y_i(0)$ represents.

(b) Contrast the meaning of $Y_i(0)$ with the meaning of Y_i.

(c) Contrast the meaning of $Y_i(0)$ with the meaning of $Y_i(1)$. Is it ever possible to observe both at the same time for unit i? Why?

(d) Explain the notation $\mathbb{E}[Y_i(0) \mid T_i = 1]$, where T_i is a binary variable that gives the treatment status for subject i, 1 if treated, 0 if control.

(e) Contrast the meaning of $\mathbb{E}[Y_i(0)]$ with the meaning of $\mathbb{E}[Y_i \mid T_i = 0]$.

(f) Contrast the meaning of $\mathbb{E}[Y_i(0) \mid T_i = 1]$ with the meaning of $\mathbb{E}[Y_i(0) \mid T_i = 0]$.

(g) Which of the following four expectations (that you explained in the previous questions) can be identified from the observed data?

Note: Identification means that an expression can be reduced to a function that contains only observed variables and no potential outcomes. Do not make any additional assumptions about the distributions of the variables, except that there is at least one observation with $T_i = 1$, and at least one with $T_i = 0$ in the observed data.

$$\mathbb{E}[Y_i(0) \mid T_i = 1]$$
$$\mathbb{E}[Y_i(0)]$$
$$\mathbb{E}[Y_i \mid T_i = 0]$$
$$\mathbb{E}[Y_i(0) \mid T_i = 0]$$

Problem 4 (10 points)

Consider the following population of six individuals, with respective treatment assignments and potential outcomes represented in the table below:

i	T_i	Y_i	$Y_i(1)$	$Y_i(0)$
1	1	4	4	3
2	0	4	8	4
3	1	2	2	0
4	1	3	3	0
5	0	8	7	8
6	0	7	6	7

Answer the following questions about this population.

(a) If you naively compare *observed* differences between treatment and control groups (difference in means estimator, $\tau_D iM$), what is your estimate of the average treatment effect? In other words, compute the value of:

$$\mathbb{E}[Y_i|T_i = 1] - \mathbb{E}[Y_i|T_i = 0]$$

(b) What is the value of selection bias with respect to $Y_i(1)$ in this example? *Hint: recall the formula of the difference in means estimator with respect to Average Treatment Effect on the Control (ATC)*

(c) Calculate value of the average treatment effect. Is it the same as the naive comparison of observed means you calculated in part (a)? Explain why or why not.

Problem 5 (10 points)

Recall the strong ignorability assumption we discussed in class, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i$. Now suppose that we weaken it instead assuming that only $Y_i(0) \perp\!\!\!\perp T_i$. In this case, it is not possible to identify the ATE by naively comparing the difference in means of Y_i between treatment and control (untreated) groups. Can you identify the average treatment effect on the treated (ATT) with a naive comparison of means? Offer formal justification for your answer.

Problem 6 (30 points)

Suppose you are interested in understanding the causal relationship between incumbency (binary) and vote share (share on a scale between 0 and 1) of a candidate in local elections.

(a) Identify at least one covariate that should be included in a causal regression of vote share on incumbency status. Explain why this variable should be included?

(b) Consider a variable that measures media coverage on candidate (e.g., number of stories about candidate per week), which can be influenced by both incumbency status and the vote share of a candidate. Discuss whether you should control for this variable in your analysis. Explain the potential implications of including or excluding this variable in your model, and whether it may introduce any biases.

(c) Should you include a control variable for campaign spending in the election prior to the one being studied? Explain whether including this variable might help or hinder causal identification, and offer reasoning to support your decision.

(d) Draw a Directed Acyclic Graph (DAG) that illustrates the main relationship between incumbency status and vote share as well as relationships between these variables and the variables you discussed in parts (a)-(c) of this problem. Describe the directions (positive or negative) of causal relationships implied by your DAG.

(e) In R simulate a dataset with 1000 observations that reflects the DAG structure you presented in part (d), including potential variables such as campaign spending, electoral success, other relevant covariates you described in part (a), and variables discussed in parts (b)-(c).

(f) Using `lm()` function in R estimate the base regression model of incumbency status on the vote share. Next, run the regression that you think will perform the best in terms of getting estimates close to the true effect (that you assumed when simulating the data). Finally run regression that includes media coverage variable and campaign spending in the election prior (parts (b) and (c)). Briefly discuss differences or similarities in estimates between campaign spending and the vote share you observe across these models.

(g) (BONUS - 10 points) Which regression model will provide estimates of the relationship between incumbency and vote share that are close to the true relationship you assumed but also have highest precision of those estimates (i.e. lowest standard error). Show this regression model output and briefly explain why you think the standard errors will be the lowest in this case.