Midterm Practice Problem Set - POL201.01

Version 2 - 03/10/2025

DISCLAIMER: This practice problem set serves as a representative sample of the types of questions that will appear on the actual exam. However, since the midterm is limited to approximately 90 minutes, it will be significantly shorter than this practice version. The goal of this practice problem set is to give you ample opportunity to reinforce your understanding and refine your problem-solving skills.
NOTE: I will post the solutions to the problem set questions Monday 03/10 (pm). So, it is strongly recommended that you start working on the questions well before the solutions are posted. Also, just reading through the solutions without trying to attempt the problems first is a poor studying strategy. Always try to solve the problems first, then use the solutions to check your work.

Section 1: Introduction to Data

Instructions: For each statement below, circle `True` or `False`, then provide a brief justification explaining your answer.

Study Context: A team of researchers surveyed a sample of farmers across various regions to understand their financial habits, experiences with government loans for implementing green economy technologies, and attitudes toward sustainable agricultural practices. The researchers did not assign any treatments or intervene in any way; they simply collected self-reported data from participants over a single time period.

Consider the first few rows of the collected sample dataset:

Business ID	Owner Age	Revenue	Loan Approved	Years in Business	Region	Business Type
1	52	180000	Yes	12	West	Retail
2	37	95000	No	5	East	Service
3	45	200000	Yes	15	South	Manufacturing
4	40	120000	No	8	South	Retail
5	50	170000	Yes	10	West	Technology
⋮	⋮	⋮	⋮	⋮	⋮	⋮

1. **Statement A:** Owner Age is a continuous numerical variable, while Business Type is an ordinal categorical variable.
   `True` `False`

2. **Statement B:** The dataset above represents an *experimental* study capable of establishing causation.
   `True` `False`

3. **Statement C:** Assume each member of the population had the same probability of being selected in the sample data. *Therefore*, the results from this sample can be generalized to the larger population of small business owners.

   `True` `False`

4. **Statement D:** If Loan Approval and Years in Business are strongly associated in the dataset, that guarantees that having more years in business causes higher loan approval rates.

   `True` `False`

5. **Statement E:** Suppose that researchers collected this data only using online surveys sent to businesses that had previously interacted with financial institutions. If the sampling was random within this group, there is no chance of sampling bias.

   `True` `False`

Section 2: Describing and Summarizing Data

In political science, analyzing the frequency of discussions about political events can provide insights into civic engagement and public opinion dynamics. Below is a dataset representing the number of times a random sample of 18 high school students engaged in political discussions with friends over the past month.

$$\{1, 1, 1, 1, 1,\ \ 2, 2, 2, 2,\ \ 3, 3, 3,\ \ 4, 4,\ \ 5,\ \ 6, 6,\ \ 7\}.$$

For this dataset, the following aggregates or summaries are provided:

$$\sum x_i = 54, \quad \sum (x_i - \bar{x})^2 = 64, \quad n = 18$$

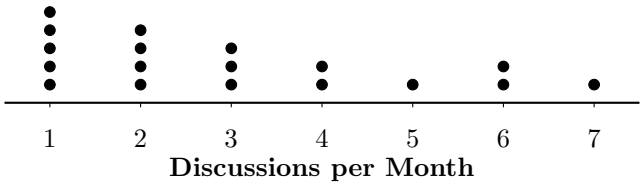

Figure 1: Dot Plot of Students' Discussions per Month (X)

Questions

Thoroughly justify all your answers.

1. **Measures of Central Tendency:**
   (a) Calculate the *mean* number of political discussions per student.
   (b) Find the *median* number of discussions.
   (c) Identify the *mode*.
   (d) Compare and interpret these measures and comment on any qualitative differences you observe.

2. **Measures of Spread:**

   (a) Compute the *range* of the data.

   (b) Determine the *interquartile range* (IQR).

   (c) Calculate the sample *standard deviation*.

   (d) Compare and interpret these measures and comment on any qualitative differences you observe.

3. **Skewness and Outliers:**

   (a) What can you infer about the skewness of the data by comparing the mean and median? Is your answer consistent with a visual inspection of the shape of the distribution?

   (b) Are there any data points you might consider as outliers in this set? Justify your answer using the $\pm 1.5 \cdot IQR$ rule of thumb to classify outliers.

4. **Impact of New Data:** Suppose a new student joins the class, reporting 15 political discussions per month.

   (a) Predict qualitatively how adding this data point would affect the *mean* and *median*.

   (b) Between the mean and the median, which measure of central tendency is more robust when a new extreme value is added? Explain why.

Section 3: Probability Rules

Instructions: For each of the following problems, write down the correct probability rule you applied and show your calculations.

1. **Question A:** A political survey finds that 40% of voters support Candidate A, 35% support Candidate B, and 25% are undecided. If one voter is selected at random, what is the probability that they support either Candidate A or Candidate B?

2. **Question B:** A school district surveyed 2,000 randomly selected students from two high schools: 1,200 from School A and 800 from School B. Of these, 500 students from School A and 400 students from School B participated in at least one extracurricular activity.
   Is the following statement `True` or `False` (justify)? "*A randomly selected student from School A is more likely to participate in at least one extracurricular activity than a randomly selected student from School B.*"

3. **Question C:** A university study finds that 20% of students do not own a laptop, and among all students. Additionally, the study finds that 48% own both a laptop and a tablet. If a student is randomly selected and found to own a laptop, what is the probability that they also own a tablet?

4. **Question D:** Suppose the probability that a randomly selected registered voter will turn out to vote in an election is 0.65. If two registered voters are randomly selected, what is the probability that both of them will vote?

5. **Question E:** A research study finds that 50% of students take a statistics course, 40% take a political science course, and 20% take both. What is the probability that a randomly selected student takes either a statistics or a political science course?

6. **Question F:** A survey shows that 30% of people get their news from social media, 50% from television, and 20% from newspapers. The probability of misinformation being shared is 0.25 for social media, 0.10 for television, and 0.05 for newspapers. What is the overall probability that a randomly chosen person encounters misinformation?

7. **Question G:** An AI tool is designed to detect misinformation in online articles. It analyzes an article and predicts whether it contains false information. The tool has the following performance characteristics:

   - If an article truly contains misinformation, the AI correctly flags it 95% of the time (true positive rate).
   - If an article is actually truthful, the AI incorrectly flags it as misinformation 15% of the time (false positive rate).
   - Before analyzing an article, the estimated probability that any randomly selected article contains misinformation is 0.2 (the prior probability).

   If the AI flags an article as misinformation, what is the probability that the article is actually false?

8. **Question H:** A political survey finds that 24% of voters support Candidate X, 15% support Candidate Y, and 61% are undecided. If one voter is selected at random, and given this voter is not undecided, what is the probability that they prefer Candidate X?

Section 4: Probability - Vote Miscounting Puzzle

Instructions: For each of the following problems, write down the correct probability rule you applied and show your calculations. Note: Do questions (a) and (b). The rest are given as a challenge.

1. **Question: Probability Puzzle on Vote Miscounting**

   In an election, Candidate A receives 55% of the votes (i.e., $P(A) = 0.55$), and Candidate B receives 45% of the votes (i.e., $P(B) = 0.45$). However, due to potential miscounting (M), each vote has a 10% probability of being counted incorrectly. This means:

   - A vote for A has a 10% probability of being miscounted as a vote for B (i.e., $P(M|A) = 0.10$), and a 90% probability of being counted correctly (i.e., $P(M^c|A) = 0.90$).
   - A vote for B has a 10% probability of being miscounted as a vote for A (i.e., $P(M|B) = 0.10$), and a 90% probability of being counted correctly (i.e., $P(M^c|B) = 0.90$).

   Using the appropriate probability rules, answer the following:

   (a) What is the probability that a randomly selected vote is counted correctly? *Tip: Use the law of total probability.*

   (b) What is the probability that a randomly selected vote is miscounted? *Tip: Use the complement rule.*

   (c) (**Challenge Question**) What is the probability that a randomly selected vote was truly cast for Candidate A before miscounting, denoted $P(A^*)$? *Tip: Compute $P(A^*)$ by considering both correctly counted votes for A and votes miscounted from B to A. Use the following version of the law of total probability:*

$$\begin{aligned} P(A^*) &= P(\text{Correctly Counted for A}) + P(\text{Miscounted for B}) \\ &= P(M^c \cap A) + P(M \cap B) = P(M^c|A)P(A) + P(M|B)P(B). \end{aligned}$$

(d) (**Challenge Question**) What is the probability that a randomly selected vote was truly cast for Candidate B before miscounting, denoted $P(B^*)$? *Tip: Compute $P(B^*)$ by considering both correctly counted votes for B and votes miscounted from A to B. Use the following version of the law of total probability:*

$$\begin{aligned} P(B^*) &= P(\text{Correctly Counted for B}) + P(\text{Miscounted for A}) \\ &= P(M^c \cap B) + P(M \cap A) = P(M^c|B)P(B) + P(M|A)P(A). \end{aligned}$$

(e) (**Challenge Question**) Given your calculations of $P(A^*)$ and $P(B^*)$, does vote miscounting change the expected winner of the election? *Tip: Compare $P(A^*)$ and $P(B^*)$ to determine if the winner remains the same.*

(f) (**Challenge Question**) Suppose instead that the probability of miscounting is some unknown value θ, where $P(M|A) = P(M|B) = \theta$. What is the largest value of θ for which Candidate A remains the expected winner? *Hint:*

(i) Express $P(A^*)$ and $P(B^*)$ in terms of θ.

(ii) Solve for θ when $P(A^*) = P(B^*)$.

(iii) Interpret the result: What does this tell you about the sensitivity of the election outcome to miscounting?

Section 5: Linear Combination of Random Variables

Instructions: For each of the following problems, apply the appropriate statistical concepts related to linear transformations of random variables, expectations, variances, and probability distributions. Clearly show all steps in your calculations.

1. **Question: Sales Commissions and Expected Earnings**

   Alex works as a salesperson for two different companies:

   - **Job A:** Sales revenue in a given week is represented by the discrete random variable X, with expectation $E[X] = 5000$ and variance $V(X) = 1,000,000$.
   - **Job B:** Sales revenue in a given week is represented by the discrete random variable Y, with expectation $E[Y] = 3000$ and variance $V(Y) = 400,000$.
   - Alex receives a **commission** of **5%** on sales from **Job A** and **10%** on sales from **Job B**.

   Alex's total commission earnings are given by:

   $$Z = 0.05X + 0.10Y$$

   Answer the following:

   (a) Compute the **expected total commission earnings** $E[Z]$.

   (b) Compute the **variance of total commission earnings** $V(Z)$, assuming that X and Y are independent.

   (c) Compute the **standard deviation of** Z, denoted σ_Z.

(d) **In terms of standard deviation units, how far is $Z = 630.62$ from the mean?**

- *Hint: Compute the **Z-score**:*

$$Z_{\text{score}}(630.62) = \frac{630.62 - E[Z]}{\sigma_Z}$$

(e) **Now assume $Z_{\textbf{score}}$ follows a standard normal distribution. Compute the probability that Alex earns more than \$630.62 in commissions, i.e., find $P(Z_{\textbf{score}} > Z_{\textbf{score}}(630.62))$.**

- *Hint: Use the **standard normal CDF table** below to find the cumulative probability $P(Z_{score} \leq Z_{score}(630.62))$.*

Z_{score}	-3	-2	-1	0	1	2	3
$\text{CDF}_Z(z)$	0.0013	0.0228	0.1587	0.5000	0.8413	0.9772	0.9987

- *Hint: Apply the **complement rule**:*

$$P(Z_{\text{score}} > Z_{\text{score}}(630.62)) = 1 - P(Z_{\text{score}} \leq Z_{\text{score}}(630.62))$$

Section 6: Random Variables II

As a government analyst, you are tasked with estimating the cost of a new food stamp program. The program has fixed administrative costs, and each participant receives a monthly benefit. The fixed administrative cost is \$6,000,000, and the monthly benefit per participant is \$200. *The number of participants is represented by a random variable P.* Below is the probability distribution for the number of participants:

Participants: P	Probability: $\Pr(P = p)$
$P < 50,000$	0
50,000	0.08
75,000	0.22
100,000	0.31
125,000	0.24
150,000	0.15
$P > 150,000$	0

Table 1: Probability Mass Function of Participants.

- **Fixed Administrative Cost:** \$6,000,000.
- **Monthly Benefit per Participant:** \$200.

Answer the following questions. Clearly show all steps in your calculations.

(a) **Calculate the Expected Value of the Number of Participants**
Determine the expected value of the number of participants in the program. *Justify your answer.*

(b) **Define the Total Monthly Cost as a Linear Combination**
Write down the total monthly cost of the program, C, as a linear combination of P, the monthly benefit per participant, and the fixed administrative cost. *Justify your answer.*

(c) **Compute the Probability Mass Function of the Total Monthly Cost**
Compute the probability mass function of the total monthly cost, C. For each possible value of the total cost, indicate its probability, $P_C(c)$. Express this in a table. *Justify your answer.*

(d) **Compute the Expected Value of the Total Monthly Cost**
Calculate the expected value of the total monthly cost of the program. *Justify your answer.*

(e) **Compute the Cumulative Distribution Function of the Total Monthly Cost**
Compute the cumulative distribution function of the total monthly cost, C. For each possible value of the total cost, calculate $P_C(C \leq c)$. Express this in a table. *Justify your answer.*

(f) **Assess Budget Constraints**
The government has allocated a maximum monthly budget of \$30,000,000 for the program. Calculate the probability that the cost of the program exceeds this budget. *Justify your answer.*