Cumulative Practice Problem Set

DS 1000B Winter 2026

Notes

This practice problem set is intended to support your preparation for your midterm and exam. The questions reflect the range of difficulty you may encounter on the actual exam; some may be more challenging, while others may be less so.

- Some problems may be approached differently depending on the chosen methodology. A key example is the analysis of skewness: you may choose to either exclude outliers before describing the tails (Marieke's approach) or include the outliers in your description (Pavel's approach).

  Neither method is objectively superior to the other; they are simply different conventions. Consequently, both approaches will be accepted for full credit.

  Students whose answers align with Marieke's framework should not be concerned if their steps differ from the text below. Grading will account for both philosophical approaches.

- These problems are **not** representative of the overall length or coverage of the midterm or final exam.
- The midterm information sheet posted by February 14th, which will include exam logistics, the provided formula sheet, and general details about exam weighting.
- You may discuss these problems during office hours, on Piazza or with your classmates.

Chapter 1: Data Visualization

Circle one answer per question.

MC1. A survey involving 35 questions was given to a sample of 200 students attending a university with an enrollment of 10,000. How many individuals do the data describe?

A. 35 **B.** **200** **C.** 10,000 **D.** It is impossible to say.

MC2. A survey involving 35 questions was given to a sample of 200 students attending a university with an enrollment of 10,000. How many variables do the data contain?

A. **35** **B.** 200 **C.** 35 × 200 **D.** 35 × 10,000

MC3. A dataset of exam scores shows a histogram with a long tail to the right. Which statement is most likely true?

A. The mean is less than the median

B. The distribution is symmetric

C. **The mean is greater than the median**

D. The mode equals the mean

E. The distribution is uniform

MC4. A stemplot of test scores is provided below:

Stem	Leaf
3	5 7 8
4	2 6 7 8 9
5	8
6	1 5 5 8 9
7	0 2 2 4 6 7
8	1 3 4
9	0 2

Key: 5—8 = 58. How many students scored in the 70s?

A. 2

B. 3

C. 4

D. 5

E. **6**

MC5. Which statement is correct regarding the distribution of a variable?

A. The distribution of a quantitative variable can be displayed using a graphical display like a histogram.

B. The distribution of a categorical variable can be displayed using a table of counts or percents.

C. The distribution of a variable tells us what values the variable takes and how often it takes these values.

D. **All of these choices are correct.**

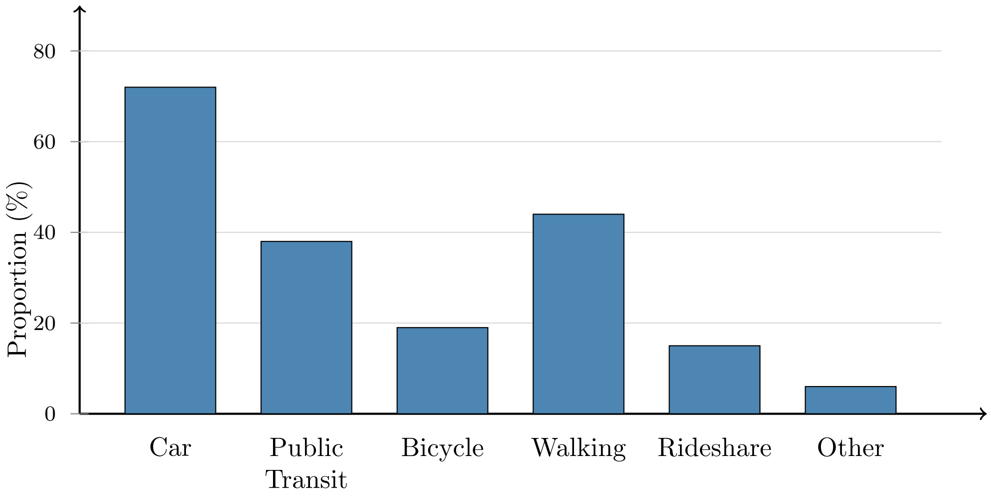
MC6. The bar chart below displays the percentage of commuters by their primary modes of transportation. Is the total proportion of people who commute by bicycle or by walking greater than, less than, or equal to the proportion who commute by car?

A. Greater than

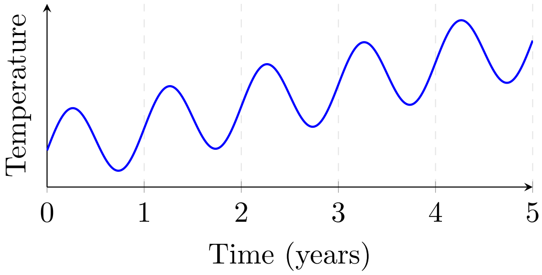
B. Equal to

C. **Less than**

D. Cannot be determined



MC7. Where does the following time series plot exhibit deviation?



A. 0

B. Approximately at 1.4

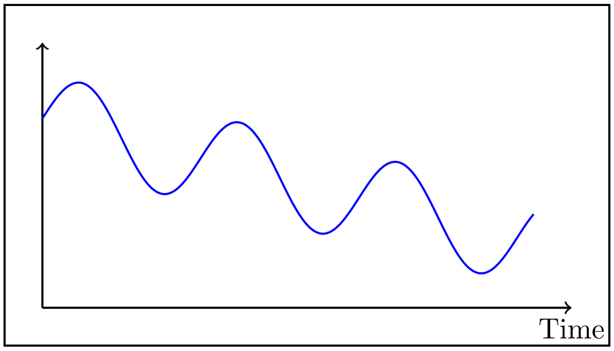
C. **There are no deviations in this plot**

D. 5

E. 2.7

Short Answer

(a) Comment on the following time series plot. What features does it exhibit?



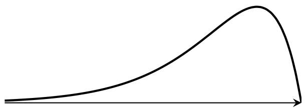
Solution:

The time series exhibits two distinct features:

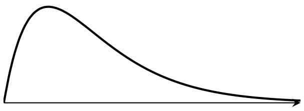
- Trend: There is a clear downward linear trend over time.
- Seasonality: There is a recurring cyclic pattern superimposed on the trend.

(b) Two density plots are sketched below. One shows the distribution of age at death from natural causes (heart disease, cancer, and so forth). The other shows age at death from trauma (accident, murder, suicide). Which is which, and why?

(i)



(ii)



Solution:

(i) Natural Causes: This distribution is left-skewed. Most deaths from natural causes occur at older ages (the peak is to the right), with fewer deaths occurring at younger ages.

(ii) Trauma: This distribution is right-skewed. Deaths from trauma (accidents, etc.) are more common among younger populations (peak is to the left) and decrease in frequency as age increases, though they can happen at any age (long right tail).

Chapter 2: Mean, Median, Quartiles, Boxplots, and Standard Deviation

MC8. A dataset has values: 10, 15, 20, 25, 100. If the observation 100 is removed, which measure of center changes the most?

A. Median **B.** Mode **C.** Mean **D.** All change equally

MC9. Two classes have the same mean test score of 75. Class A has a standard deviation of 5, while Class B has a standard deviation of 15. Which statement is true?

A. Class A scores are more spread out than Class B

B. **Class B scores are more spread out than Class A**

C. Both classes have identical score distributions

D. Class A has higher scores on average

E. Standard deviation does not measure spread

MC10. The five number summary of a dataset is given by: Min = 20, Q1 = 30, Median = 40, Q3 = 50, Max = 60. What is the interquartile range (IQR)?

A. 10 **B.** 30 **C.** **20** **D.** 40

MC11. A dataset has mean 50 and standard deviation 10. If every value is multiplied by 2 and then 5 is added, what are the new mean and standard deviation?

A. Mean = 50, SD = 10

B. Mean = 100, SD = 20

C. **Mean = 105, SD = 20**

D. Mean = 105, SD = 10

E. Mean = 55, SD = 25

MC12. For a right-skewed distribution, which inequality is most likely true?

A. Mean ¡ Median ¡ Mode

B. **Mode ¡ Median ¡ Mean**

C. Median ¡ Mode ¡ Mean

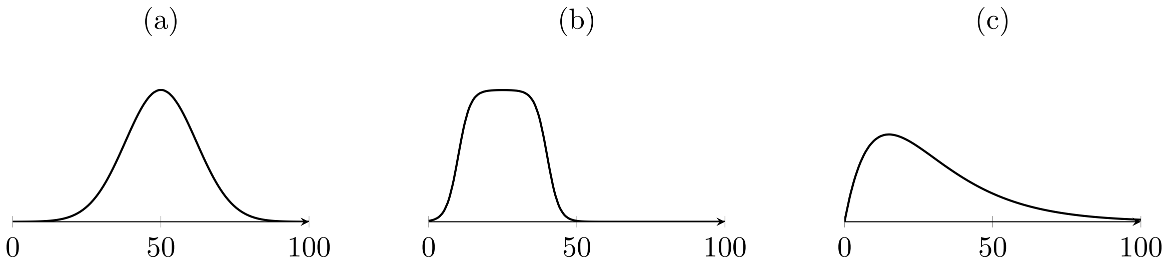
D. Mean = Median = Mode

E. Cannot be determined without data

Short Answer

(a) For each density plot below, indicate if the median is greater than (>), less than (<), or approximately equal to ($\approx$) the mean.

Solution: (a) Symmetric: Mean $\approx$ Median.
(b) Relatively symmetric on the left side: Mean $\approx$ Median.
(c) Right-skewed: Mean $>$ Median.



(b) For each density plot above, approximate the mode of the distribution.

a) Mode: ________ b) Mode: ________ c) Mode: ________

Solution: (a) Mode $\approx$ 50 (peak of symmetric distribution).
(b) Mode $\approx$ 25 (flat top/plateau centered around 25).
(c) Mode $\approx$ 15 (peak of right-skewed distribution).

(c) Compute the mean and median of the following data: 2, 4, 6, 8, 10.

Solution: Mean: $(2 + 4 + 6 + 8 + 10)/5 = 30/5 = 6$.

Median: The middle value (3rd of 5) is 6.

$$\boxed{\text{Mean} = 6, \ \text{Median} = 6}$$

(d) Compute the mean and median of the following data: 1, 3, 3, 6, 7, 8, 9.

Solution: Mean: $(1 + 3 + 3 + 6 + 7 + 8 + 9)/7 = 37/7 \approx 5.29$.

Median: The middle value (4th of 7) is 6.

$$\boxed{\text{Mean} \approx 5.29, \ \text{Median} = 6}$$

(e) Find the third quartile (Q_3) of the following data: 3, 5, 7, 9, 12, 15, 18.

Solution: Ordered data: 3, 5, 7, 9, 12, 15, 18 ($n = 7$).

Q_3 is the median of the upper half (values above the median). Median is the 4th value: 9.

Upper half: 12, 15, 18.

Median of upper half: 15

$$\boxed{Q_3 = 15}$$

Question 2

A dataset consists of the values 10, 15, 20, 25, and a fifth unknown value x. After including x, the median of the dataset is 18.

What are the possible values of x? Explain your reasoning.

Solution: Arrange the five values in increasing order. The median is the middle value after sorting.

For the median to be 18, x must be placed so that 18 is the third value. Consider all possible positions for x:

- If $x \leq 10$: Sorted order is x, 10, 15, 20, 25. Median is 15.
- If $10 < x \leq 15$: Sorted order is 10, x, 15, 20, 25. Median is 15.
- If $15 < x < 20$: Sorted order is 10, 15, x, 20, 25. Median is x.
- If $x = 20$: Sorted order is 10, 15, 20, 20, 25. Median is 20.
- If $20 < x < 25$: Sorted order is 10, 15, 20, x, 25. Median is 20.
- If $x \geq 25$: Sorted order is 10, 15, 20, 25, x. Median is 20.

Therefore, for the median to be 18, x must satisfy $15 < x < 20$ and $x = 18$.

$$\boxed{x = 18}$$

Chapter 4: Scatterplots and Correlation

MC13. A researcher finds that the correlation between hours of sleep and reaction time is $r = -0.72$. If a new dataset is collected where reaction times are measured in milliseconds instead of seconds, what happens to the correlation coefficient?

A. It becomes positive

B. It increases in magnitude

C. **It stays the same**

D. It becomes zero

E. It decreases in magnitude

MC14. A study finds a correlation of $r = 0.89$ between ice cream sales and drowning deaths. Which conclusion is most appropriate?

A. Ice cream consumption causes drowning

B. Drowning causes people to buy ice cream

C. **A third variable (such as hot weather) may explain both**

D. The correlation proves a causal relationship

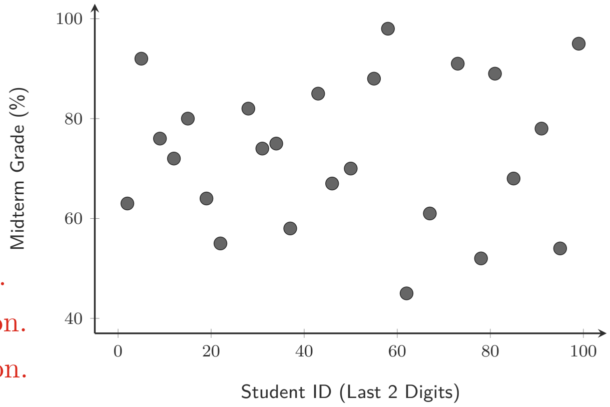
E. The relationship must be nonlinear

Short Answer

(a) What is a scatterplot?

Solution: A scatterplot is a graphical representation that displays the relationship between two quantitative variables. Each point on the scatterplot represents an observation, with its horizontal position determined by one variable and its vertical position determined by the other.

(b) Which features do we examine when commenting on a scatterplot? Comment on them for the scatterplot below.



Outliers, form, direction, strength

- Outliers: there appear to be no outliers.
- Form: there appears to be no association.
- Direction: N/A as there is no association.
- Strength: N/A as there is no association.

Solution:

Chapter 5: Linear Regression

MC15. In a regression predicting test scores from study hours, the equation is $\hat{y} = 45 + 5x$. A student who studies 8 hours has an actual score of 82. What is this student's residual?

A. 37 **B.** 40 **C.** 74 **D.** -3 **E.** 3

Solution: Predicted score: $45 + 5 \times 8 = 85$. Residual $= 82 - 85 = -3$.

MC16. Suppose the variables x and y have standard deviations $s_x = s_y = 1$. What is the slope of the regression line when regressing y on x?

A. The slope is 0

B. The slope is 1

C. **The slope equals the correlation coefficient r**

D. The slope is s_y/s_x

E. Cannot be determined without knowing the means

Solution: When $s_x = s_y$, the slope formula $b = r\frac{s_y}{s_x}$ simplifies to $b = r\frac{1}{1} = r$.

$$\boxed{b = r}$$

Question 3

A marketing analyst studies the relationship between advertising spending (in thousands of dollars) and monthly sales (in thousands of units). Data from 15 months yields the estimated regression:

$$\widehat{\text{Sales}} = 12.5 + 3.2 \times \text{Advertising}$$

The correlation coefficient is $r = 0.85$.

(a) Interpret the slope and intercept in context of the problem.

Solution: Slope (3.2): For each additional \$1,000 spent on advertising, sales increase by approximately 3,200 units on average.

Intercept (12.5): When advertising spending is \$0, the model predicts sales of 12,500 units.

(b) Calculate R^2 and interpret it in context of the problem.

Solution: $R^2 = r^2 = 0.85^2 = 0.7225$.

Approximately 72.25% of the variability in monthly sales is explained by advertising spending.

$$\boxed{R^2 = 72.25\%}$$

(c) One month had advertising spending of $8,000 and actual sales of 42,000 units. Calculate the residual. Is this an overestimate or underestimate?

Solution: Predicted: $\hat{y} = 12.5 + 3.2(8) = 12.5 + 25.6 = 38.1$ (thousand units).

Residual $= y - \hat{y} = 42 - 38.1 = 3.9$ (thousand units).

Positive residual means the model underestimated actual sales.

$$\boxed{\text{Residual} = +3,900 \text{ units (underestimate)}}$$

Question 4

The table below shows the relationship between hours of study (x) and test scores (y) for four students:

Hours of Study (x)	Test Score (y)
1	50
2	70
3	65
4	85

(a) Determine the least squares regression line for predicting test score from hours of study.

Solution: Step 1: Calculate means.

$$\bar{x} = \frac{1+2+3+4}{4} = 2.5, \qquad \bar{y} = \frac{50+70+65+85}{4} = 67.5$$

Step 2: Calculate standard deviations.

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{5}{3}} \approx 1.291$$

$$s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{(-17.5)^2 + (2.5)^2 + (-2.5)^2 + (17.5)^2}{3}} = \sqrt{\frac{625}{3}} \approx 14.434$$

Step 3: Calculate standardized values and correlation.

x	y	$z_x = \frac{x-\bar{x}}{s_x}$	$z_y = \frac{y-\bar{y}}{s_y}$	$z_x \cdot z_y$
1	50	-1.161	-1.213	1.408
2	70	-0.387	0.173	-0.067
3	65	0.387	-0.173	-0.067
4	85	1.161	1.213	1.408
				Sum: 2.682

$$r = \frac{\sum z_x z_y}{n-1} = \frac{2.682}{3} \approx 0.894$$

Step 4: Calculate slope and intercept.

$$b = r\frac{s_y}{s_x} = 0.894 \times \frac{14.434}{1.291} \approx 10$$

$$a = \bar{y} - b\bar{x} = 67.5 - 10(2.5) = 42.5$$

$$\boxed{\hat{y} = 42.5 + 10x}$$

(b) Interpret the slope of the regression line in the context of the problem.

Solution: For each additional hour of study, the predicted test score increases by 10 points on average.

(c) A student studies for 2.5 hours. What is the predicted test score for this student?

Solution:

$$\hat{y} = 42.5 + 10(2.5) = 42.5 + 25 = 67.5$$

$$\boxed{\hat{y} = 67.5}$$

Question 5

Suppose that we know that the regression line of y onto x is given by

$$\hat{y} = 3 + 0.4x$$

If $s_y = 2$, $s_x = 4$, $\bar{x} = 0$, find an expression for the regression line of x onto y.

Solution: From the regression of y on x: $\hat{y} = 3 + 0.4x$, the slope is $b = 0.4$.

Using $b = r\frac{s_y}{s_x}$ with $s_y = 2$ and $s_x = 4$:

$$r \cdot \frac{2}{4} = 0.4 \implies r = 0.8$$

For x on y, the slope is:

$$b' = r\frac{s_x}{s_y} = 0.8 \times \frac{4}{2} = 1.6$$

Since $\bar{x} = 0$ and the $(\bar{x}, \bar{y})$ lies on the regression line,

$$\bar{y} = 3 + 0.4(0) = 3.$$

[Substituting the point $x = \bar{x}$ into the regression line to find $\bar{y}$.]

The intercept is:

$$a' = \bar{x} - b'\bar{y} = 0 - 1.6(3) = -4.8$$

$$\boxed{\hat{x} = -4.8 + 1.6y}$$

Question 6

Reference: DS1000A Midterm Questions 23–24

Context: Suppose it was found that the least-squares regression line predicting house price (y) from size in square feet (x) is given by

$$\hat{y} = 20000 + 150x$$

Q23: Which of the following gives the least-squares regression line predicting square footage (y) based on house price (x)?

It was claimed that the correct answer is:

The line cannot be determined from the information provided.

Explain why this is the correct answer and indicate what additional information would be needed to find the regression line of square footage on house price.

Solution: The regression line of y on x gives us the slope $b = r\frac{s_y}{s_x}$, but this alone does not determine r, s_x, or s_y individually.

To find the regression line of x on y, we need the slope $b' = r\frac{s_x}{s_y}$. Since we only know the product $r\frac{s_y}{s_x} = 150$, we cannot determine $r\frac{s_x}{s_y}$ without additional information.

That is, we would need at least two of the three quantities: r, s_x, and s_y. We could then solve for the third and compute the regression line where the explanatory and response variables are reversed.

Chapter 6: Contingency Tables

Question 7

The popularity of computer, video, online, and virtual reality games has raised concerns about their ability to negatively affect youth. The data below are based on a recent survey of 14 to 18 year olds in Connecticut high schools.

	Grade Average		
	A's and B's	C's	D's and F's
Played games	736	450	193
Never played games	205	144	80

(a) How many people does this table describe? How many of these have played video games?

Solution: Total people:

$$(736 + 450 + 193) + (205 + 144 + 80) = 1379 + 429 = 1808$$

Number who played games:

$$736 + 450 + 193 = 1379$$

$$\boxed{\text{Total} = 1808, \quad \text{Played games} = 1379}$$

(b) Find the marginal distribution of grades. What percentage of students received a grade of C or lower?

Solution: First, calculate column totals:

- A's and B's: $736 + 205 = 941$
- C's: $450 + 144 = 594$
- D's and F's: $193 + 80 = 273$

Percent with C or lower:

$$\frac{594 + 273}{1808} = \frac{867}{1808} \approx 48.0\%$$

$$\boxed{48.0\%}$$

(c) Determine the conditional distribution of grades for those who play video games.

Solution: Total who played games: 1379.

Conditional Distribution:

- A's and B's: $736/1379 \approx 53.4\%$
- C's: $450/1379 \approx 32.6\%$
- D's and F's: $193/1379 \approx 14.0\%$

(d) Find the relative risk of receiving a grade of C or lower for those who play video games compared to those who do not.

Solution: Calculate probabilities:

$$P(\text{C or lower} \mid \text{Played}) = \frac{450+193}{1379} = \frac{643}{1379} \approx 0.466$$
$$P(\text{C or lower} \mid \text{Never Played}) = \frac{144+80}{429} = \frac{224}{429} \approx 0.522$$

Relative Risk:

$$RR = \frac{P(\text{C or lower} \mid \text{Played})}{P(\text{C or lower} \mid \text{Never Played})} = \frac{0.466}{0.522} \approx 0.893$$

$$\boxed{RR \approx 0.893}$$

(e) Find the relative odds of playing video games for those who received A's and B's compared to those who received D's and F's.

Solution: Calculate odds:

$$\text{Odds(Played} \mid \text{A's and B's)} = \frac{736}{205} \approx 3.59$$
$$\text{Odds(Played} \mid \text{D's and F's)} = \frac{193}{80} \approx 2.41$$

Odds Ratio:

$$OR = \frac{3.59}{2.41} \approx 1.49$$

$$\boxed{OR \approx 1.49}$$

(f) Interpret the relative odds calculated in part (e) in the context of this problem.

Solution: The odds ratio of approximately 1.49 indicates that the odds of having played video games are about 1.49 times higher for students who received A's and B's compared to students who received D's and F's. This suggests that students with higher grades are more likely to have played video games than students with lower grades.

Chapters 12 and 13: Probability

MC17. A weighted four-sided die has the following probabilities: 1 shows with probability x/c, 2 with $2x/c$, 3 with $4y/c$, and 4 with $8z/c$, where $x, y, z > 0$. Which of the following must be true of c?

A. $c = x + 2x + 4y + 8z$

B. $c = x + y + z$

C. $c = 2x + 4y + 8z$

D. $c = 3x + 4y + 8z$

Solution: For a valid probability distribution, probabilities must sum to 1:

$$\frac{x}{c} + \frac{2x}{c} + \frac{4y}{c} + \frac{8z}{c} = 1 \implies \frac{3x + 4y + 8z}{c} = 1 \implies c = 3x + 4y + 8z$$

Short Answer

(a) What are the two types of random variables? Explain the difference(s) between them.

Solution: There are two types of random variables: discrete and continuous.

Discrete random variables take on a countable number of distinct values. Examples include the number of heads in coin flips.

Continuous random variables can take on any value within a given range. Examples include height, weight, or temperature.

Question 8

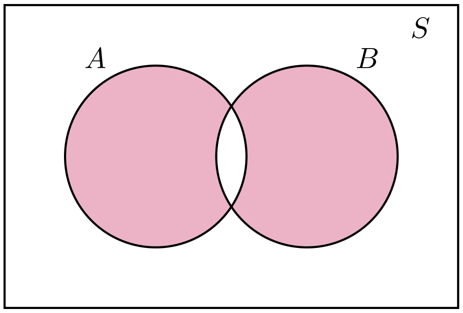
The symmetric difference of two events A and B is defined as $A\Delta B = (A \cap B^c) \cup (B \cap A^c)$. Let S be the outcome space consisting of all people. Suppose that A and B are the events

A: people who like apples.

B: people who like bananas.

(a) Draw a Venn diagram to illustrate the event $A\Delta B$.

Solution:



The shaded region represents $A\Delta B$: elements in A only or in B only, but not in both.

(b) Interpret the symmetric difference $A\Delta B$ in the context of this problem.

Solution: The symmetric difference $A\Delta B$ represents people who like either apples or bananas, but not both. It includes people who like only apples and people who like only bananas.

(c) Explain why $P(A\Delta B) = P(A \cup B) - P(A \cap B)$.

Solution: The symmetric difference $A\Delta B$ consists of everything in the union $A \cup B$ except the intersection $A \cap B$. Since $A \cap B$ is a subset of $A \cup B$, we have:

$$A\Delta B = (A \cup B) \setminus (A \cap B)$$

Therefore, $P(A\Delta B) = P(A \cup B) - P(A \cap B)$.

(d) If $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$, calculate $P(A\Delta B)$.

Solution:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.4 + 0.5 - 0.2 = 0.7$$
$$P(A\Delta B) = P(A \cup B) - P(A \cap B) = 0.7 - 0.2 = 0.5$$

$$\boxed{P(A\Delta B) = 0.5}$$

(e) Suppose A and B are independent and that $P(A) = 0.5$ and $P(B^c) = 0.1$. Compute $P(A\Delta B)$.

Solution: Given $P(A) = 0.5$ and $P(B^c) = 0.1$, so $P(B) = 0.9$.

Since A and B are independent:

$$P(A \cap B) = P(A)P(B) = 0.5 \times 0.9 = 0.45$$

Using independence:

$$P(A \cap B^c) = P(A)P(B^c) = 0.5 \times 0.1 = 0.05$$

$$P(A^c \cap B) = P(A^c)P(B) = 0.5 \times 0.9 = 0.45$$

$$P(A\Delta B) = P(A \cap B^c) + P(A^c \cap B) = 0.05 + 0.45 = 0.5$$

$$\boxed{P(A\Delta B) = 0.5}$$

Question 9

A coin is tossed 3 times and the result of each toss is recorded as H (heads) or T (tails).

(a) What is an outcome of this random phenomenon?

Solution: An outcome is a specific sequence of heads and tails from the 3 tosses, e.g., HHT, TTH, etc.

(b) List the sample space S.

Solution: $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$

(c) Define the event A as "at most 1 tail observed." List the outcomes in event A.

Solution: $A = \{HHH, HHT, HTH, THH\}$

(d) What is the probability of event A?

Solution: $P(A) = \frac{|A|}{|S|} = \frac{4}{8} = 0.5$

$$\boxed{P(A) = 0.5}$$

(e) Let B be the event "each face appears at least once." List the outcomes in event B.

Solution: $B = \{HHT, HTH, HTT, THH, THT, TTH\}$

(f) Are events A and B independent? Justify your answer.

Solution: $P(A) = 4/8 = 0.5$ and $P(B) = 6/8 = 0.75$.

$A \cap B = \{HHT, HTH, THH\}$, so $P(A \cap B) = 3/8 = 0.375$.

Since $P(A) \times P(B) = 0.5 \times 0.75 = 0.375 = P(A \cap B)$, events A and B are independent.

$$\boxed{\text{Yes, } A \text{ and } B \text{ are independent.}}$$

(g) Are the events A^c and B^c independent? Justify your answer.

Solution: $A^c = \{HTT, THT, TTH, TTT\}$, so $P(A^c) = 4/8 = 0.5$.

$B^c = \{HHH, TTT\}$, so $P(B^c) = 2/8 = 0.25$.

$A^c \cap B^c = \{TTT\}$, so $P(A^c \cap B^c) = 1/8 = 0.125$.

Since $P(A^c) \times P(B^c) = 0.5 \times 0.25 = 0.125 = P(A^c \cap B^c)$, events A^c and B^c are independent.

Note: If A and B are independent, then A^c and B^c are also independent.

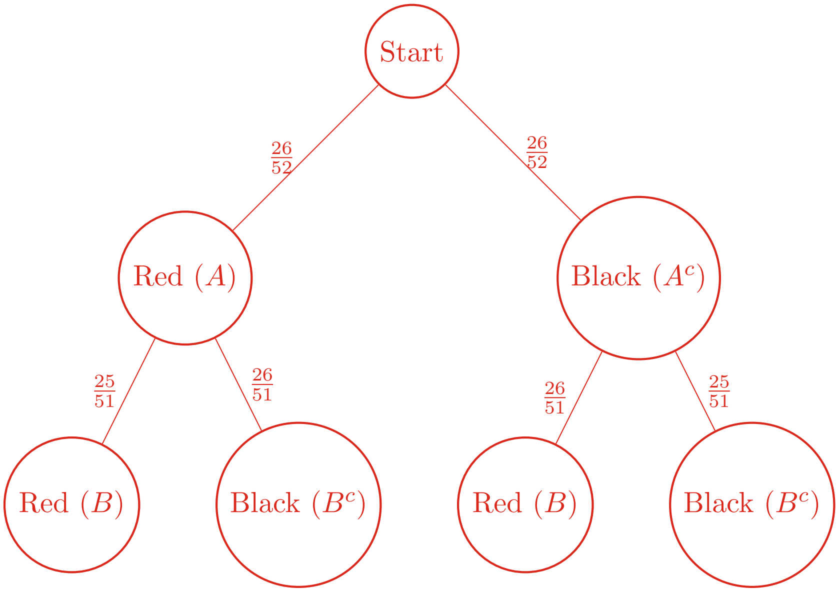
$$\boxed{\text{Yes, } A^c \text{ and } B^c \text{ are independent.}}$$

Question 10

Two cards are drawn without replacement from a standard deck of 52 playing cards. Let A be the event that the first card drawn is red, and let B be the event that the second card drawn is red.

(a) Draw a tree diagram representing this scenario, labeling all branches with their probabilities.

Solution:



(b) Are events A and B independent? Justify your answer.

Solution: No, A and B are not independent. Drawing without replacement means the outcome of the first draw affects the composition of the deck for the second draw.

Alternatively: $P(B \mid A) = 25/51 \neq P(B) = 1/2$, so they are not independent.

(c) Calculate $P(A \cap B)$.

Solution:

$$P(A \cap B) = P(A) \times P(B \mid A) = \frac{26}{52} \times \frac{25}{51} = \frac{1}{2} \times \frac{25}{51} = \frac{25}{102}$$

$$\boxed{P(A \cap B) = \frac{25}{102}}$$

(d) Compute $P(B \mid A)$ and $P(A \mid B)$. Compare the two values.

Solution:

$$P(B \mid A) = \frac{25}{51} \approx 0.4902$$

For $P(A \mid B)$, use Bayes' theorem:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{25/102}{1/2} = \frac{25}{51} \approx 0.4902$$

Both probabilities are equal: $P(B \mid A) = P(A \mid B) = \frac{25}{51}$.

$$\boxed{P(B \mid A) = P(A \mid B) = \frac{25}{51}}$$

(e) Compute $P(A)$ and $P(B)$. Compare the two values.

Solution: $P(A) = 26/52 = 1/2$.

Using the law of total probability for $P(B)$:

$$\begin{aligned} P(B) &= P(B \mid A)P(A) + P(B \mid A^c)P(A^c) \\ &= \frac{25}{51} \cdot \frac{1}{2} + \frac{26}{51} \cdot \frac{1}{2} = \frac{25+26}{51 \cdot 2} = \frac{51}{102} = \frac{1}{2} \end{aligned}$$

$P(A) = P(B) = \frac{1}{2}$.

$$\boxed{P(A) = P(B) = \frac{1}{2}}$$

Question 11

In a match between Brazil and Germany, let X, Y denote the number of goals scored by Brazil and Germany, respectively. The distributions are:

x	0	1	2	3	4	5	6	7
$P(X = x)$	0	c	0.2	0.2	0.2	0.2	0	0

y	0	1	2	3	4	5	6	7
$P(Y = y)$	$2c$	c	c	$c/2$	0	0	0	d

Assume X and Y are independent.

(a) Find values of c and d that make X and Y valid random variables.

Solution: For X: $c + 0.2 + 0.2 + 0.2 + 0.2 = 1 \implies c = 0.2$.

For Y: $2c + c + c + c/2 + d = 1 \implies 4.5(0.2) + d = 1 \implies 0.9 + d = 1 \implies d = 0.1$.

$$\boxed{c = 0.2, \quad d = 0.1}$$

(b) What is the probability that Germany scores 7 goals and Brazil scores 1 goal?

Solution: By independence:

$$P(X = 1, Y = 7) = P(X = 1) \cdot P(Y = 7) = 0.2 \times 0.1 = 0.02$$

$$\boxed{P(X = 1, Y = 7) = 0.02}$$

(c) Determine the probability the game ends in a tie.

Solution: A tie occurs when $X = Y$. Using the distributions with $c = 0.2$ and $d = 0.1$:

$$\begin{aligned} P(\text{Tie}) &= P(X=1, Y=1) + P(X=2, Y=2) + P(X=3, Y=3) \\ &= (0.2)(0.2) + (0.2)(0.2) + (0.2)(0.1) \\ &= 0.04 + 0.04 + 0.02 = 0.10 \end{aligned}$$

$$\boxed{P(\text{Tie}) = 0.10}$$

(d) Determine the probability that Germany wins given that Brazil scored at least 4 goals.

Solution:

$$P(\text{Germany wins} \mid X \geq 4) = \frac{P(Y > X \text{ and } X \geq 4)}{P(X \geq 4)}$$

First, $P(X \geq 4) = P(X = 4) + P(X = 5) = 0.2 + 0.2 = 0.4$.

Germany wins when $Y > X$:

Case 1: $X = 4$. Germany wins if $Y > 4$, only $P(Y = 7) = 0.1$ is nonzero.

$P(X = 4, Y = 7) = 0.2 \times 0.1 = 0.02$

Case 2: $X = 5$. Germany wins if $Y > 5$, only $P(Y = 7) = 0.1$ is nonzero.

$P(X = 5, Y = 7) = 0.2 \times 0.1 = 0.02$

Total numerator: $0.02 + 0.02 = 0.04$

$$P(\text{Germany wins} \mid X \geq 4) = \frac{0.04}{0.4} = 0.1$$

$$\boxed{0.1}$$

Question 12

During press conferences, President Donald Trump is sometimes asked questions about geopolitics. When this occurs, there is a 50% chance that he goes on a tangent. He derides Joe Biden or boasts about having the "hottest economy" only if he is on a tangent.

- If he goes on a tangent, there is an 80% chance that he mentions Joe Biden.
- If he mentions Joe Biden, there is a 40% chance that he also mentions the "hottest economy."
- The overall probability that he mentions the "hottest economy" is 25%.

Let T, B, H be the events defined as follows

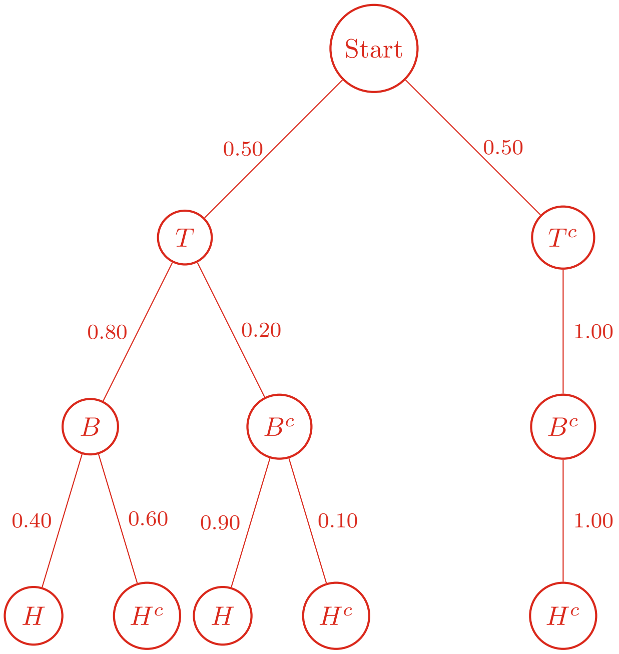
T: Trump goes on a tangent

B: Trump derides Joe Biden

H: Trump boasts about having the "hottest economy"

(a) Draw a tree diagram representing this scenario.

Solution:



(b) Find the probability $P(B \cup T)$.

Solution: Since Biden is only mentioned during tangents, $B \subseteq T$, which means $B \cap T = B$.

Therefore:

$$P(B \cup T) = P(T) = 0.50$$

$$\boxed{P(B \cup T) = 0.50}$$

(c) Find the probability $P(T \cap B \cap H)$.

Solution:

$$\begin{aligned} P(T \cap B \cap H) &= P(T) \cdot P(B \mid T) \cdot P(H \mid T \cap B) \\ &= 0.50 \times 0.80 \times 0.40 \\ &= 0.16 \end{aligned}$$

$$\boxed{P(T \cap B \cap H) = 0.16}$$

(d) Find the probability that Trump boasts about having the "hottest economy" given that he did not deride Joe Biden.

Solution: We need to find $P(H \mid B^c)$.

First, calculate $P(B)$. Since Biden is only mentioned during tangents:

$$P(B) = P(B \mid T)P(T) = 0.80 \times 0.50 = 0.40$$

Therefore, $P(B^c) = 1 - 0.40 = 0.60$.

Next, find $P(H \cap B)$:

$$P(H \cap B) = P(H \mid B)P(B) = 0.40 \times 0.40 = 0.16$$

Therefore,

$$\begin{aligned} P(H) &= P(H \cap B) + P(H \cap B^c) \\ 0.25 &= 0.16 + P(H \cap B^c) \\ P(H \cap B^c) &= 0.09 \end{aligned}$$

Thus,

$$P(H \mid B^c) = \frac{P(H \cap B^c)}{P(B^c)} = \frac{0.09}{0.60} = 0.15$$

$$\boxed{P(H \mid B^c) = 0.15}$$

(e) Find the probability $P(H \cup T)$.

Solution: Since H only occurs during tangents, $H \subseteq T$, so $H \cap T = H$.

Therefore:

$$\begin{aligned} P(H \cup T) &= P(H) + P(T) - P(H \cap T) \\ &= 0.25 + 0.50 - 0.25 \\ &= 0.50 \end{aligned}$$

$$\boxed{P(H \cup T) = 0.50}$$

(f) Determine $P(T \cap B^c \cap H^c)$, the probability that Trump goes on a tangent but mentions neither Joe Biden nor the "hottest economy."

Solution: We use the multiplication rule for dependent events:

$$P(T \cap B^c \cap H^c) = P(T) \cdot P(B^c \mid T) \cdot P(H^c \mid T \cap B^c)$$

Given:

$$\begin{aligned} P(T) &= 0.50 \\ P(B^c \mid T) &= 1 - P(B \mid T) = 1 - 0.80 = 0.20 \\ P(H^c \mid T \cap B^c) &= 1 - P(H \mid T \cap B^c) = 1 - 0.90 = 0.10 \end{aligned}$$

Therefore,

$$P(T \cap B^c \cap H^c) = 0.50 \times 0.20 \times 0.10 = 0.01$$

$$\boxed{P(T \cap B^c \cap H^c) = 0.01}$$

Chapter 3: Normal Distribution

Question 13

Argentina is testing javelin throwers for the Olympics. Distances are normally distributed with mean 70 meters and standard deviation 5 meters.

(a) What is the probability an athlete breaks the Olympic record of 92.97 meters?

Solution: Calculate the z-score:

$$z = \frac{92.97 - 70}{5} = \frac{22.97}{5} \approx 4.59$$

The probability that an athlete throws farther than 92.97 meters is

$$P(X > 92.97) = P(Z > 4.59) \approx \boxed{0}$$

(b) Find x such that 99% of throws do not land farther than x.

Solution: We need the 99th percentile. From the standard normal table, $z_{0.99} \approx 2.326$.

$$x = \mu + z \cdot \sigma = 70 + 2.326(5) = 70 + 11.63 = \boxed{81.63 \text{ m}}$$

(c) Athletes are shortlisted if their throw exceeds 80 meters. Given that an athlete was shortlisted, what is the probability their throw was at most 85 meters?

Solution: First find $P(X > 80)$: $z = (80 - 70)/5 = 2$, so $P(X > 80) = P(Z > 2) \approx 0.0228$.

Next find $P(X > 85)$: $z = (85 - 70)/5 = 3$, so $P(X > 85) = P(Z > 3) \approx 0.00135$.

Then $P(80 < X \leq 85) = 0.0228 - 0.00135 = 0.02145$.

Therefore:

$$P(X \leq 85 \mid X > 80) = \frac{P(80 < X \leq 85)}{P(X > 80)} = \frac{0.02145}{0.0228} \approx 0.94$$

$$\boxed{P(X \leq 85 \mid X > 80) \approx 0.94}$$

Chapter 15: Sampling Distributions

Question 14

A farm ships crates of potatoes. Potato weight has mean $\mu = 300$g and standard deviation $\sigma = 50$g. Each crate contains 100 potatoes. An inspector randomly samples crates and calculates the average weight per potato $\bar{X}$.

(a) What are the mean and standard deviation of the average potato weight per crate $\bar{X}$?

Solution: Mean: $\mu_{\bar{X}} = \mu = 300$g.

Standard deviation (standard error): $\sigma_{\bar{X}} = \sigma/\sqrt{n} = 50/\sqrt{100} = 5$g.

$$\boxed{\mu_{\bar{X}} = 300\text{g}, \quad \sigma_{\bar{X}} = 5\text{g}}$$

(b) Approximate the probability that the average weight per potato in a crate is between 295g and 305g.

Solution: Standardize: $z_1 = (295 - 300)/5 = -1$ and $z_2 = (305 - 300)/5 = 1$.

By the empirical rule:

$$P(-1 < Z < 1) \approx 0.68$$

$$\boxed{P \approx 68\%}$$

(c) If the inspector wants to be 95% confident that the average weight per potato in a sampled crate is within 3g of the true mean, how many potatoes should they sample?

Solution: For 95% confidence, $z^* = 1.96$. We need:

$$1.96 \cdot \frac{50}{\sqrt{n}} \leq 3 \implies \sqrt{n} \geq \frac{1.96 \times 50}{3} \approx 32.67 \implies n \geq 1067.3$$

$$\boxed{n = 1068 \text{ potatoes}}$$

(d) Find the probability that the total weight of potatoes in a crate exceeds 31 kg.

Solution: Let $S = \sum_{i=1}^{100} X_i$ be the total weight. Then $S = 100\bar{X}$.

We want $P(S > 31{,}000\text{g}) = P(\bar{X} > 310\text{g})$.

Compute the z-score:

$$z = \frac{310 - 300}{5} = 2$$

Therefore:

$$P(\bar{X} > 310) = P(Z > 2) \approx 0.0228$$

$$\boxed{\approx 0.0228}$$

Chapter 16: Confidence Intervals

Question 15

A battery manufacturer wants to estimate the average lifetime of their AA batteries. The population standard deviation is known to be $\sigma = 12$ hours. A random sample of 49 batteries has a sample mean lifetime of 98 hours.

(a) Calculate the standard error of the sample mean.

Solution:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{49}} = \frac{12}{7} \approx 1.71 \text{ hours}$$

$$\boxed{SE \approx 1.71 \text{ hours}}$$

(b) Construct a 99% confidence interval for the mean battery lifetime.

Solution: For 99% confidence, $z^* = 2.576$.

$$98 \pm 2.576 \times 1.71 \approx 98 \pm 4.41 \implies (93.59, 102.41)$$

$$\boxed{(93.59, 102.41) \text{ hours}}$$

(c) How many batteries must be sampled to achieve a 95% confidence interval with margin of error no more than 2 hours?

Solution: For 95% confidence, $z^* = 1.96$. We need:

$$1.96 \cdot \frac{12}{\sqrt{n}} \leq 2 \implies \sqrt{n} \geq \frac{1.96 \cdot 12}{2} = 11.76 \implies n \geq 138.3$$

$$\boxed{n = 139 \text{ batteries}}$$

Question 16

A random sample of 64 light bulbs has a mean lifetime of $\overline{x} = 1200$ hours. Assume the population standard deviation is $\sigma = 150$ hours.

(a) Construct a 90% confidence interval for the mean lifetime of all bulbs.

Solution: For 90% confidence, $z^* = 1.645$. Standard error: $SE = 150/\sqrt{64} = 18.75$.

$$1200 \pm 1.645 \times 18.75 = 1200 \pm 30.84 = \boxed{(1169.16,\ 1230.84)}$$

(b) Interpret your interval from part (a).

Solution: We are 90% confident that the true mean lifetime of all light bulbs lies between 1169.16 hours and 1230.84 hours.

(c) Recall that the total width of a confidence interval is the distance between the upper and lower bounds. If the interval has total width 40 hours and the sample size is still 64, what is the confidence level?

Solution: Width $= 2z^* \times SE = 2z^* \times 18.75 = 40$

Solve for z^*: $z^* = \frac{40}{2\times18.75} = \frac{40}{37.5} \approx 1.067$

$\Phi(1.067) \approx 0.857$

Confidence level $= 2 \times 0.857 - 1 = 0.714$ or about 71.4%.

$$\boxed{\approx 71.4\%}$$

Chapters 7 and 8: Sampling and Experiments

Question 17

A researcher is interested in learning the proportion of university students that use recreational drugs. They decide to survey students who visit the campus library at opening. A simple random sample of 50 students is taken, and each student is asked whether they use drugs recreationally.

(a) Define the population, sample, parameter, and statistic in this study.

Solution: Population: All university students.

Sample: The 50 students surveyed at the campus library.

Parameter: The true proportion of university students who use recreational drugs.

Statistic: The proportion of surveyed students who report using recreational drugs.

(b) Identify two potential sources of bias in this sampling process.

Solution: Possible biases include:

1. Selection bias: Only students who visit the library at opening are sampled, systematically excluding those who don't visit the library or visit at other times.

2. Response bias: Students who use recreational drugs may be less likely to respond truthfully due to social desirability concerns.

(c) Suppose the population standard deviation is $\sigma = 0.5$. Calculate the standard error for a sample of $n = 50$ students.

Solution:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{50}} \approx 0.071$$

$$\boxed{SE \approx 0.071}$$

(d) Define stratified random sampling.

Solution: Stratified random sampling involves dividing the population into distinct subgroups (strata) that share a characteristic, then randomly sampling from each stratum.

For this study, strata could be defined by year of study (e.g., first, second, third, fourth year) or by faculty/major.

Question 18

A high school wants to investigate whether using flashcards improves students' math exam scores compared to traditional study methods.

(a) Identify the subjects, factor, and treatments in this experiment.

Solution: Subjects: The students participating in the study.

Factor: The study method.

Treatments: (1) Using flashcards, (2) Using traditional study methods.

(b) Describe a block design that could reduce variance in this study.

Solution: Group students into blocks based on prior math performance (e.g., low, medium, high GPA). Within each block, randomly assign students to flashcards or traditional study. This controls for prior ability.

(c) Identify two potential lurking variables.

Solution: 1. Motivation level: More motivated students may choose flashcards and also perform better independently.

2. Study time: Students using one method may study longer than those using the other.

(d) Define double blinding and explain its importance. Is it feasible in this experiment?

Solution: Double blinding is a technique in experimental design where neither the participants nor the experimenters know which treatment the participants are receiving. This helps prevent bias in treatment administration and outcome assessment.

In this experiment, double blinding is not feasible because students will know which study method they are using.