Do NOT write in the area above this line.

**Practice Final Exam**

**Course**: DS 1000B Winter 2026

**Sections**: 002, 003

**Date**:

180 minutes

**Instructors**:

Marieke Mur

Pavel Shuldiner

**Allowed aids**:

A calculator (non-programmable, non-graphing)

Formula sheet (provided).

Standard Normal Table (provided).

| | |
|---|---|
| Full Name (print)<br>*(e.g. Tom Marvolo Riddle):* | |
| Western ID<br>*(e.g. baldemort13):* | |
| Student Number<br>*(e.g. 251123456):* | |

1. Legibly **print** your Western User ID, full name, and student number in the spaces provided above.
2. Do **not** detach the pages of the exam. You may ask for scrap paper if needed.
3. The space at the top of each page is reserved for the scanner. Do not write on or near the barcode.
4. The exam has 24 pages. The last three pages may be used for additional workspace or scrap paper.

| **Section** | **Marks** |
|---|---|
| Multiple Choice (1 mark each) | 10 |
| Exploring Data | 6 |
| Study Hours | 10 |
| Height and Shoe Size | 10 |
| Exercise and Sleep Quality | 13 |
| Fair and Biased Coins | 11 |
| Wine Bottle Volumes | 17 |
| Reverse Curving Grades | 12 |
| Student Preferences and Activities | 8 |
| Educational Pathways | 8 |
| **Total** | **106** |

## Multiple Choice (1 mark each) **(10 marks)**

Each of the following multiple-choice questions is worth 1 mark. Circle the correct answer for each question. No justification is required and no partial marks will be awarded.

**MC1.** (1 mark) A regression model predicting final exam score from hours studied has $R^2 = 0.64$. The best interpretation is:

**A.** **64% of the variability in exam scores is explained by the linear relationship with hours studied.**

**B.** The correlation between hours studied and exam score is 0.64.

**C.** 64% of students passed the exam.

**D.** For each additional hour studied, the exam score increases by 64%.

**E.** 64% of students studied enough hours to do well.

**MC2.** (1 mark) A data set has a mean of 50 and a median of 48. If the largest value in the data set is replaced by a value ten times as large, which of the following is true?

**A.** Both the mean and median increase.

**B.** The median increases but the mean stays the same.

**C.** Both the mean and median stay the same.

**D.** Both the mean and median decrease.

**E.** **The mean increases but the median stays the same.**

**MC3.** (1 mark) A histogram of household incomes in a city is strongly right-skewed. Which of the following is most likely true?

**A.** The mean is less than the median.

**B.** The mean equals the median.

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

**D.** The standard deviation equals zero.

**E.** The distribution is symmetric.

**MC4.** (1 mark) A random variable $X$ takes values 0, 1, 2, and 3 with probabilities 0.10, 0.25, 0.40, and 0.25 respectively. What is $P(X \geq 2)$?

**A.** 0.25 **B.** 0.40 **C.** **0.65** **D.** 0.75 **E.** 0.90

**MC5.** (1 mark) A researcher measures the heights of 50 randomly selected students and calculates a mean of 170 cm. The value 170 cm is best described as:

**A.** A population

**B.** **A statistic**

**C.** A variable

**D.** A sample space

**E.** A parameter

**MC6.** (1 mark) In a regression analysis, a data point has a residual of $-5$. This means:

**A.** The observed value is 5 units above the predicted value.

**B.** The slope of the regression line is $-5$.

**C.** **The predicted value is 5 units above the observed value.**

**D.** The correlation coefficient is $-5$.

**E.** The regression model has an error of 5%.

**MC7.** (1 mark) As the number of coin flips increases, the proportion of heads observed tends to get closer to 0.5. This phenomenon is explained by:

**A.** The Central Limit Theorem

**B.** Bayes' theorem

**C.** The empirical rule

**D.** **The Law of Large Numbers**

**E.** The addition rule

**MC8.** (1 mark) Which of the following best describes what the standard deviation measures?

**A.** The middle value of the data set.

**B.** The difference between the largest and smallest values.

**C.** The most frequently occurring value.

**D.** The average of all data values.

**E.** **The typical distance of data values from the mean.**

**MC9.** (1 mark) A data set has $Q_1 = 20$ and $Q_3 = 40$. Using the $1.5 \times$ IQR rule, a value is considered an outlier if it is:

**A.** Below 10 or above 50.

**B.** Below 0 or above 60.

**C.** Below 5 or above 55.

**D.** **Below $-10$ or above 70.**

**E.** Below 15 or above 45.

**MC10.** (1 mark) You interview 10 randomly selected workers about their commute distance and compute the sample mean. If you repeat this survey many times, recording each sample mean, what does a histogram of these sample means represent?

**A.** The bias present in the sampling method

**B.** The true population mean commute distance

**C.** A simple random sample

**D.** **The sampling distribution of the sample mean**

**E.** The confidence interval for the population mean

## Exploring Data **(6 marks)**

**Q1.** A student council conducts a survey of 200 students about their study habits and preferences. Answer the following questions based on their findings.

(a) (2 marks) The survey found that 35% of students prefer studying at the library, 30% at home, 25% at a coffee shop, and 20% in study rooms. Can a pie chart be drawn to display these results? Explain.

*Solution:* No, a pie chart cannot be drawn. The percentages sum to $35 + 30 + 25 + 20 = 110\%$, which exceeds 100%. This suggests students could select more than one location. A pie chart requires mutually exclusive categories whose proportions sum to exactly 100%.

(b) (2 marks) The survey also recorded each student's weekly study hours. Is this variable categorical or quantitative? Name one appropriate visualization for displaying the distribution of this variable.

*Solution:* The variable is quantitative. An appropriate visualization is a histogram, which shows the shape, centre, and spread of the distribution. One could also use a stemplot to display the distribution of study hours.

Note that a modified boxplot would not be a good visualization for displaying the distribution as it only captures a few of the features of the distribution (quartiles, outliers) but does not provide much on the granularity of the data.

(c) (2 marks) What does correlation measure?

*Solution:* Correlation measures the strength and direction of the linear relationship between two quantitative variables.

## Study Hours **(10 marks)**

**Q2.** The following data represent the number of hours studied per week by 12 randomly selected university students:

8, 13, 15, 17, 18, 19, 21, 22, 24, 26, 28, 30

(a) (2 marks) Draw a stemplot for this data.

*Solution:*

```
0 | 8
1 | 3 5 7 8 9
2 | 1 2 4 6 8
3 | 0
```

Note: Each stem represents tens, each leaf represents ones.

(b) (1 mark) Identify the shape of the distribution.

*Solution:* The distribution is approximately symmetric.

(c) (5 marks) Calculate the five-number summary for this data.

*Solution:* The five-number summary is:

- Minimum: 8
- First Quartile (Q1): 16
- Median (Q2): 20
- Third Quartile (Q3): 25
- Maximum: 30

(d) (2 marks) Based on the shape you identified in part (b), which measures of center and spread would be most appropriate to summarize this data? Justify your answer.

*Solution:* Since the distribution is approximately symmetric, the mean and standard deviation are the most appropriate measures of centre and spread.

## Height and Shoe Size **(10 marks)**

**Q3.** A researcher studies the relationship between height (in inches) and shoe size (US men's sizing) among college students. After collecting data, they obtain the following least squares regression line:

$$\widehat{\text{Height}} = 50 + 2.5 \times (\text{Shoe Size})$$

(a) (2 marks) Suppose the residual for a particular observation is 1 inch and the observed height is 74 inches. What is the shoe size of this student? Show your work.

*Solution:* The residual is the difference between observed and predicted height:

$$\begin{aligned} \text{Residual} &= \text{Observed} - \text{Predicted} \\ 1 &= 74 - \text{Predicted} \\ \text{Predicted height} &= 73 \text{ inches} \end{aligned}$$

Using the regression equation:

$$\begin{aligned} 73 &= 50 + 2.5 \times (\text{Shoe Size}) \\ 23 &= 2.5 \times (\text{Shoe Size}) \\ \text{Shoe Size} &= \frac{23}{2.5} = \boxed{9.2} \end{aligned}$$

(b) (2 marks) If the mean height of the sampled students is 66.67 inches, what is the mean shoe size? Show your work.

*Solution:* The regression line passes through $(\bar{x}, \bar{y})$, where $x$ is shoe size and $y$ is height.
Given $\overline{y} = 66.67$ inches,

$$\begin{aligned} 66.67 &= 50 + 2.5 \times \overline{x} \\ 16.67 &= 2.5 \times \overline{x} \\ \overline{x} &= \frac{16.67}{2.5} = \boxed{6.67} \end{aligned}$$

(c) (3 marks) Using the same data, the researchers found that the least squares regression line for predicting shoe size from height is

$$\widehat{\text{Shoe Size}} = -4.0 + 0.16 \times (\text{Height})$$

Calculate the correlation coefficient ($r$) between height and shoe size. Show your work.

*Solution:* Let $b$ be the slope for predicting height from shoe size, and $b'$ be the slope for predicting shoe size from height.

Note that since $b = r \times \frac{s_y}{s_x}$ and $b' = r \times \frac{s_x}{s_y}$, multiplying these slopes together gives

$$bb' = r \times \frac{s_y}{s_x} \times r \times \frac{s_x}{s_y} = r^2$$

Therefore, the correlation coefficient is a square root of $bb'$. Since both slopes are positive, the correlation must also be positive.

$$\begin{aligned} r &= \text{sign}(b) \times \sqrt{b \times b'} \\ &= (+1) \times \sqrt{2.5 \times 0.16} \\ &= \sqrt{0.4} \\ &= 0.632 \end{aligned}$$

$$\boxed{r = 0.63}$$

(d) (3 marks) If we measure height in centimetres instead of inches (where 1 inch = 2.54 cm), would the correlation coefficient change? Would the slope of the regression line change? Explain.

*Solution:* **Correlation coefficient:** No, $r$ would **not** change. Correlation is a unitless measure of the strength and direction of a linear relationship. It is unaffected by linear transformations of the variables.

**Slope:** Yes, the slope **would** change. Converting height from inches to centimetres multiplies by 2.54, which scales the standard deviation of height by the same factor, resulting in a different slope:

$$a_{\text{new}} = 2.5 \times 2.54 = 6.35 \quad \text{(for predicting cm-height from shoe size)}$$

## Exercise and Sleep Quality **(13 marks)**

**Q4.** The table below shows data from a survey about exercise habits and sleep quality among 500 adults.

| | **Poor** | **Fair** | **Good** | **Total** |
|---|---|---|---|---|
| Exercises regularly | 30 | $x$ | 160 | |
| Does not exercise regularly | $y$ | 100 | $z$ | |
| Total | | | | 500 |

where $x$, $y$, and $z$ represent unknown cell counts:

- $x$ = number of adults who exercise regularly and report "Fair" sleep
- $y$ = number of adults who do not exercise regularly and report "Poor" sleep
- $z$ = number of adults who do not exercise regularly and report "Good" sleep

(a) (2 marks) Suppose it is known that 200 of the adults surveyed exercise regularly. Find $x$.

*Solution:* The row total for "Exercises regularly" is 200:

$$30 + x + 160 = 200 \implies x = 200 - 30 - 160 = \boxed{10}$$

(b) (4 marks) The relative risk of having good sleep quality for those who exercise regularly versus those who do not is 2. Determine $z$, the number of adults who do not exercise regularly and report "Good" sleep quality. Show your work.

*Solution:* Since 200 exercise and 500 total, 300 do not exercise.

$$\text{RR} = \frac{P(\text{Good} \mid \text{Exercise})}{P(\text{Good} \mid \text{No Exercise})} = 2$$

The proportion with good sleep among exercisers is:

$$P(\text{Good} \mid \text{Exercise}) = \frac{160}{200} = 0.80$$

Therefore:

$$\frac{0.80}{z/300} = 2$$
$$z/300 = 0.40$$
$$z = \boxed{120}$$

(c) (3 marks) Calculate the odds ratio of having poor sleep quality for those who do not exercise regularly compared to those who exercise regularly. Show your work.

*Solution:* First find $y$: the row total for non-exercisers is 300, so $y + 100 + 120 = 300$, giving $y = 80$.

Odds of poor sleep for non-exercisers:

$$\frac{80}{100 + 120} = \frac{80}{220}$$

Odds of poor sleep for exercisers:

$$\frac{30}{10+160} = \frac{30}{170}$$

$$\text{OR} = \frac{80/220}{30/170} = \frac{80 \times 170}{220 \times 30} = \frac{13{,}600}{6{,}600} \approx \boxed{2.06}$$

(d) (2 marks) Interpret the odds ratio from part (c) in the context of this problem.

*Solution:* The odds ratio of approximately 2.06 means that the odds of having poor sleep quality are about twice as high for adults who do not exercise regularly compared to those who exercise regularly.

(e) (2 marks) Is this study an experiment or an observational study? Can we conclude that exercise causes better sleep quality? Explain.

*Solution:* This is an **observational study** because the researchers did not assign exercise habits to participants; they simply surveyed existing behaviour.

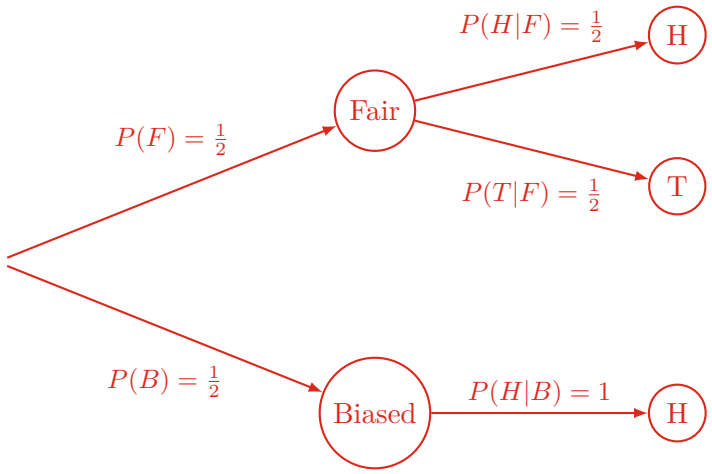
We **cannot** conclude that exercise causes better sleep. There may be confounding variables (e.g., age, stress levels, diet) that affect both exercise habits and sleep quality.

## Fair and Biased Coins **(11 marks)**

**Q5.** A bag contains two coins: one is a fair coin, and the other is a biased coin with heads on both sides. One coin is selected at random from the bag, flipped once, and the result is observed.

(a) (2 marks) Draw a tree diagram to represent all possible outcomes of this experiment. Label all branches with their probabilities.

*Solution:*




The sample space can be expressed as $S = \{(Fair, H), (Fair, T), (Biased, H)\}$.

(b) (1 mark) What is the probability of selecting the fair coin?

*Solution:* Since there are two coins and one is selected at random:

$$\boxed{P(\text{Fair}) = \frac{1}{2}}$$

(c) (1 mark) Given that the biased coin was selected, what is the probability of observing heads?

*Solution:* The biased coin has heads on both sides, so

$$\boxed{P(H \mid \text{Biased}) = 1}$$

(d) (2 marks) What is the probability of observing heads?

*Solution:* By the Law of Total Probability:

$$\begin{aligned} P(H) &= P(\text{Fair}) \times P(H \mid \text{Fair}) + P(\text{Biased}) \times P(H \mid \text{Biased}) \\ &= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times 1 \\ &= \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \end{aligned}$$

$$\boxed{P(H) = \frac{3}{4}}$$

(e) (3 marks) Given that heads was observed, what is the probability that the biased coin was selected? Show your work.

*Solution:* By Bayes' rule

$$\begin{aligned} P(\text{Biased} \mid H) &= \frac{P(\text{Biased}) \times P(H \mid \text{Biased})}{P(H)} \\ &= \frac{\frac{1}{2} \times 1}{\frac{3}{4}} \\ &= \frac{1/2}{3/4} = \frac{2}{3} \end{aligned}$$

$$\boxed{P(\text{Biased} \mid H) = \frac{2}{3}}$$

(f) (2 marks) What is the probability that either tails is observed or the biased coin is selected?

*Solution:* By the addition rule:

$$\begin{aligned} P(T \text{ or Biased}) &= P(T) + P(\text{Biased}) - P(T \text{ and Biased}) \\ &= \frac{1}{4} + \frac{1}{2} - 0 \\ &= \frac{3}{4} \end{aligned}$$

Note: $P(T \text{ and Biased}) = 0$ since the biased coin cannot show tails.

$$\boxed{P(T \text{ or Biased}) = \frac{3}{4}}$$

## Wine Bottle Volumes **(17 marks)**

**Q6.** The volume of wine in a set of bottles is known to follow a normal distribution with standard deviation $\sigma = 10$ mL.

(a) (5 marks) A random sample of 25 bottles is taken, and the sample mean volume is $\overline{x} = 750$ mL. Construct a 95% confidence interval for the true mean volume $\mu$. Show your work.

*Solution:* Standard error: $\text{SE} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$ mL

For 95% confidence: $z^* = 1.96$

$$\overline{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}} = 750 \pm (1.96)(2)$$
$$= 750 \pm 3.92$$

$$\boxed{(746.08,\ 753.92) \text{ mL}}$$

(b) (5 marks) For the same data, construct a 99% confidence interval. Compare its width[1] to the 95% interval and explain the difference.

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

$$\overline{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}} = 750 \pm (2.576)(2)$$
$$= 750 \pm 5.15$$

$$\boxed{(744.85,\ 755.15) \text{ mL}}$$

The 99% interval (width = 10.30 mL) is wider than the 95% interval (width = 7.84 mL) because higher confidence requires a larger margin of error to ensure we capture the true parameter with greater certainty.

[1] The width of a confidence interval is the difference between its upper and lower bounds.

(c) (7 marks) How many bottles must be sampled to obtain a 95% confidence interval for $\mu$ with total width equal to 1 mL? Show your work.

*Solution:* Total width of 95% CI: Width $= 2 \times z^* \times \frac{\sigma}{\sqrt{n}}$

Setting width equal to 1 mL with $z^* = 1.96$ and $\sigma = 10$:

$$\begin{aligned} 2 \times 1.96 \times \frac{10}{\sqrt{n}} &= 1 \\ \frac{39.2}{\sqrt{n}} &= 1 \\ \sqrt{n} &= 39.2 \\ n &= 1536.64 \end{aligned}$$

Rounding up, we find

$$\boxed{n = 1{,}537 \text{ bottles}}$$

## Reverse Curving Grades **(12 marks)**

**Q7.** A statistics professor's midterm was too easy. The raw scores for 200 students are normally distributed with mean $\mu = 82$ and standard deviation $\sigma = 8$. The department requires that the class average align with a target of 70. The professor decides to apply a linear "reverse curve" of the form

$$Y = aX + b$$

where $X$ is the unadjusted score and $Y$ is the adjusted score.

(a) (2 marks) What proportion of students had an unadjusted score above 90? Show your work.

*Solution:*

$$z = \frac{90 - 82}{8} = 1.00$$

From the standard normal table: $P(Z > 1.00) = 1 - 0.8413 = 0.1587$.

$$\boxed{P(X > 90) \approx 0.16 \text{ (about 16\%)}}$$

(b) (2 marks) Find the 50th percentile of the unadjusted score distribution.

*Solution:* The 50th percentile corresponds to $z = 0$ from the standard normal table.

$$x = \mu + z\sigma = 82 + (0)(8) = 82$$

(c) (3 marks) If the standard deviation of the curved scores is 12, find the value of $a$. Show your work.

*Solution:* If $Y = aX + b$ and $X \sim N(82, 8)$, then

- SD of $Y$: $|a| \cdot \sigma = 8|a| = 12 \implies |a| = 1.5$

Since $a < 0$ would reverse the ranking of students, we take $a > 0$, so $a = 1.5$.

(d) (2 marks) Using the requirement that the curved class average is 70, find the value of $b$. Show your work.

*Solution:* Comment: this part of the question is slightly outside the scope of this offering of DS1000.

Approach 1: Linearity of expectation

The mean of the curved scores must equal 70, so $E(Y) = 70$ and hence

$$\begin{aligned} E(Y) &= a \cdot \mu + b \\ 70 &= 1.5(82) + b \\ 70 &= 123 + b \\ b &= 70 - 123 = \boxed{-53} \end{aligned}$$

Approach 2: Properties of the normal distribution

Recall that standardizing a normal random variable results in a $N(0, 1)$ random variable. Hence,

$$\begin{aligned} Z &= \frac{X - \mu_X}{\sigma_X} \\ &= \frac{Y - \mu_Y}{\sigma_Y} \\ &= \frac{aX + b - \mu_Y}{|a|\,\sigma_X} \\ &= \frac{a}{|a|} \cdot \frac{X - \dfrac{\mu_Y - b}{a}}{\sigma_X} \\ &= \frac{X - \dfrac{\mu_Y - b}{a}}{\sigma_X} \end{aligned}$$

So, $\mu_X = \dfrac{\mu_Y - b}{a}$. Substituting $\mu_X = 82$, $\mu_Y = 70$, and $a = 1.5$ gives the same value for $b$:

$$\begin{aligned} 82 &= \frac{70 - b}{1.5} \\ 123 &= 70 - b \\ b &= \boxed{-53} \end{aligned}$$

(e) (3 marks) What raw score does a student need to earn at least an A (80 or above) on the curved exam? Show your work.

*Solution:* Set $Y = 80$ and solve for $X$:

$$\begin{aligned} 80 &= 1.5X - 53 \\ 1.5X &= 133 \\ X &= 88.67 \end{aligned}$$

A student needs a raw score of at least 88.67 (about 89).

## Student Preferences and Activities **(8 marks)**

**Q8.** A survey is conducted among students about their preferences and activities. Consider the events $A$, $B$ and $C$ where,

$A$ : Owns a bicycle

$B$ : Prefers tea over coffee

$C$ : Member of chess club

- $A$ and $B$ are independent
- $A$ and $C$ are mutually exclusive
- $P(A) = 0.30,\ \ P(B) = 0.40,\ \ P(C) = 0.20$
- $P(B \text{ and } C) = 0.10$

(a) (2 marks) Calculate the probability that a student owns a bicycle or prefers tea over coffee. Show your work.

*Solution:* Since $A$ and $B$ are independent:

$$\begin{aligned} P(A \text{ or } B) &= P(A) + P(B) - P(A \text{ and } B) \\ &= P(A) + P(B) - P(A) \cdot P(B) \quad \text{(by independence)} \\ &= 0.30 + 0.40 - (0.30)(0.40) \\ &= 0.70 - 0.12 \\ &= 0.58 \end{aligned}$$

$$\boxed{P(A \text{ or } B) = 0.58}$$

(b) (1 mark) Determine the probability that a student owns a bicycle and is a member of the campus chess club.

*Solution:* Since $A$ and $C$ are mutually exclusive, they cannot occur together so

$$P(A \text{ and } C) = 0.$$

$$\boxed{P(A \text{ and } C) = 0}$$

Recall that a survey is conducted among students about their preferences and activities. Consider the events $A$, $B$ and $C$ where,

| | |
|---|---|
| $A$ : Owns a bicycle<br>$B$ : Prefers tea over coffee<br>$C$ : Member of chess club | • $A$ and $B$ are independent<br>• $A$ and $C$ are mutually exclusive<br>• $P(A) = 0.30,\ P(B) = 0.40,\ P(C) = 0.20$<br>• $P(B \text{ and } C) = 0.10$ |

(c) (3 marks) Calculate the probability that a student prefers tea given that they are in the chess club. Show your work.

*Solution:* Using conditional probability definition:

$$\begin{aligned} P(B|C) &= \frac{P(B \text{ and } C)}{P(C)} \\ &= \frac{0.10}{0.20} \\ &= 0.50 \end{aligned}$$

$$\boxed{P(B|C) = 0.50}$$

(d) (3 marks) Calculate the probability that a student owns a bicycle or prefers tea over coffee, or is a member of the campus chess club; that is, find $P(A \text{ or } B \text{ or } C)$.

*Hint : Use the general addition rule for three events:*

$$P(A \text{ or } B \text{ or } C) = P(A)+P(B)+P(C)-P(A \text{ and } B)-P(B \text{ and } C)-P(A \text{ and } C)+P(A \text{ and } B \text{ and } C)$$

*Solution:* We need to find each term in the formula for $P(A \text{ or } B \text{ or } C)$:

- $P(A) = 0.30$, $P(B) = 0.40$, $P(C) = 0.20$
- $P(A \text{ and } B) = 0.30 \times 0.40 = 0.12$ (independence)
- $P(B \text{ and } C) = 0.10$ (given)
- $P(A \text{ and } C) = 0$ (mutually exclusive)
- $P(A \text{ and } B \text{ and } C) = 0$ (given)

Substituting:

$$\begin{aligned} P(A \text{ or } B \text{ or } C) &= 0.30 + 0.40 + 0.20 - 0.12 - 0.10 - 0 + 0 \\ &= 0.90 - 0.22 \\ &= 0.68 \end{aligned}$$

$$\boxed{P(A \text{ or } B \text{ or } C) = 0.68}$$

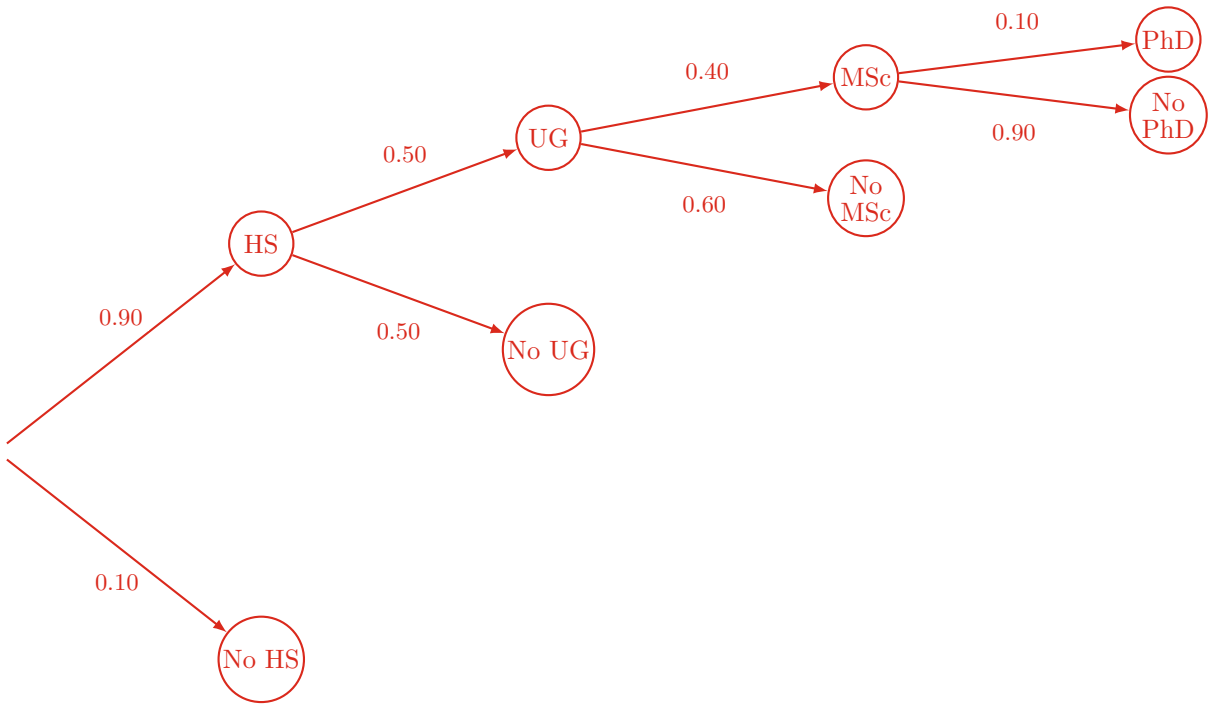
## Educational Pathways **(8 marks)**

**Q9.** Assume the educational path is sequential: finishing high school is required before starting undergraduate studies; an undergraduate degree is required before pursuing a master's degree; a master's degree is required before a Ph.D. Each stage depends on completing the previous one.

Suppose 90% of Canadian students finish high school. Of those who finish high school, 50% complete an undergraduate degree. Of those who complete undergraduate, 40% earn a master's degree. Of those who earn a master's, 10% earn a Ph.D.

(a) (2 marks) Draw a tree diagram to represent the educational pathways and label all branches with their probabilities.

*Solution:*




(b) (2 marks) What percentage of students earn a Ph.D.?

*Solution:* We multiply the probabilities at each stage:

$$\begin{aligned} P(\text{Ph.D.}) &= P(\text{HS}) \times P(\text{Undergrad}|\text{HS}) \times P(\text{Master's}|\text{Undergrad}) \times P(\text{Ph.D.}|\text{Master's}) \\ &= 0.90 \times 0.50 \times 0.40 \times 0.10 \\ &= 0.018 \end{aligned}$$

$$\boxed{P(\text{Ph.D.}) = 1.8\%}$$

(c) (2 marks) Find the probability that a student finished high school but did not earn a Ph.D.

*Solution:* We find the probability of finishing high school minus the probability of earning a Ph.D.:

$$\begin{aligned} P(\text{HS but no Ph.D.}) &= P(\text{HS}) - P(\text{Ph.D.}) \\ &= 0.90 - 0.018 \\ &= 0.882 \end{aligned}$$

$$\boxed{P(\text{HS but no Ph.D.}) = 88.2\%}$$

(d) (2 marks) Given that a student completed a Ph.D., find the probability that they earned an undergraduate degree.

*Solution:* Since earning a Ph.D. requires completing undergraduate studies (the path is sequential), every Ph.D. holder must have earned an undergraduate degree. Thus,

$$\boxed{P(\text{Undergrad} \mid \text{Ph.D.}) = 1}$$

### Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.

### Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.

### Additional Workspace

Use this space for additional work if needed. Clearly indicate which question you are working on and reference the page number.