EDEXCEL IAL MATHEMATICS - STATISTICS 1 (S1) COMPLETE STUDY GUIDE

Unit S1: Statistics 1

Assessment Overview

- **Duration:** 1 hour 30 minutes
- **Marks:** 75 marks
- **Calculator:** Permitted
- **Formulae Booklet:** Provided

Key Formulae Provided in Exam

$$\text{Mean} = \bar{x} = \frac{\sum x}{n} \text{ or } \frac{\sum fx}{\sum f}$$

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

$$\text{Interquartile range} = IQR = Q_3 - Q_1$$

$$P(A') = 1 - P(A)$$

For independent events A and B: $P(B \mid A) = P(B), P(A \mid B) = P(A), P(A \cap B) = P(A) \times P(B)$

$$E(aX + b) = aE(X) + b$$

$$\text{Var}(aX + b) = a^2\text{Var}(X)$$

Cumulative Distribution Function: $F(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} p(x)$

Standardised Normal Variable: $Z = \frac{X-\mu}{\sigma}$ where $X \sim N(\mu, \sigma^2)$

TOPIC 1: MATHEMATICAL MODELS IN PROBABILITY AND STATISTICS

1.1 Understanding Statistical Modelling

Definition: A statistical model is a simplification of a real-world situation that uses mathematics to describe, predict, or understand phenomena.

The Modelling Cycle:

1. **Observe** the real-world problem
2. **Formulate** a mathematical model
3. **Collect data** to test the model
4. **Compare** observed results with predictions
5. **Refine** the model if necessary

1.2 Types of Data

Type	Description	Examples
Qualitative	Non-numerical categories	Colour, gender, brand
Quantitative	Numerical measurements	Height, weight, score
Discrete	Countable, specific values	Number of children, dice score
Continuous	Any value in an interval	Height, time, temperature

1.3 Advantages and Disadvantages of Models

Advantages:

- Quick and inexpensive to create
- Helps understand complex real-world situations
- Enables predictions about future outcomes
- Can be refined and improved

Disadvantages:

- Simplifies reality - may miss important details
- May only work within certain ranges
- Predictions may not always be accurate

TOPIC 2: REPRESENTATION AND SUMMARY OF DATA

2.1 Stem and Leaf Diagrams

Construction:

- Stem represents the first digit(s)
- Leaves represent the final digit(s)
- Leaves must be in order
- Always include a key

Example: Data: 23, 25, 27, 31, 32, 33, 35, 42, 44, 45

```
<TEXT>
2 | 3 5 7
3 | 1 2 3 5
4 | 2 4 5
Key: 2|3 means 23
```

2.2 Back-to-Back Stem and Leaf Diagrams

Used to compare two distributions.

Example:

Group A: 23, 25, 28, 31, 33, 35 Group B: 24, 27, 29, 30, 32, 34, 36

```
<TEXT>
Group A    |    Group B
           |
3 5 3 1    |   2   4
          7    9
          0    2 4 6
Key: 3|1|0 means 31 and 30
```

Comparisons to make:

- Which group has higher median?
- Which group has larger spread?
- Are there any outliers?

2.3 Box Plots

Components:

- Minimum value
- Lower quartile (Q_1)
- Median (Q_2)
- Upper quartile (Q_3)
- Maximum value

Construction:

```
<TEXT>
|-----|-----o-----|-----|
Min    Q₁     Q₂     Q₃    Max
```

Outliers: Values outside $Q_1 - 1.5 \times IQR$to $Q_3 + 1.5 \times IQR$

Example: Draw a box plot for: 12, 15, 17, 19, 22, 25, 28, 32, 35

Solution:

- Minimum = 12, Maximum = 35
- Median (Q_2) = 22

- $Q_1 = (15 + 17)/2 = 16$
- $Q_3 = (28 + 32)/2 = 30$
- IQR = 30 - 16 = 14

2.4 Histograms

Key Points:

- Used for continuous data
- **NO gaps** between bars
- Area of bar is proportional to frequency
- Height = Frequency Density = $\frac{\text{frequency}}{\text{class width}}$

Frequency Density Formula: Frequency Density $= \frac{\text{Frequency}}{\text{Class Width}}$

Example: Draw a histogram for:

Class Interval	Frequency	Class Width	Frequency Density
$0 \leq x < 10$	15	10	1.5
$10 \leq x < 20$	25	10	2.5
$20 \leq x < 30$	35	10	3.5
$30 \leq x < 50$	40	20	2.0

TOPIC 3: MEASURES OF LOCATION AND SPREAD

3.1 Measures of Location

Mean ($\bar{x}$)

For ungrouped data: $\bar{x} = \frac{\sum x}{n}$

For grouped data (with frequencies): $\bar{x} = \frac{\sum fx}{\sum f}$

Example: Find the mean of: 4, 6, 8, 10, 12

$$\bar{x} = \frac{4 + 6 + 8 + 10 + 12}{5} = \frac{40}{5} = 8$$

Example (Grouped):

x	f	fx
2	3	6
4	5	20
6	7	42

x	f	fx
8	4	32
	∑f = 19	∑fx = 100

$$\bar{x} = \frac{100}{19} = 5.26$$

Median

For odd number of values: The middle value when ordered.

For even number of values: The mean of the two middle values.

Example: Find median of: 3, 7, 2, 9, 5 Ordered: 2, 3, 5, 7, 9 Median = 5

Mode

The most frequently occurring value.

Example: Find mode of: 4, 2, 7, 4, 3, 4, 8 Mode = 4

When to Use Each

Measure	Best Used When
Mean	Data is symmetrical, no extreme values
Median	Data is skewed, contains outliers
Mode	Data is categorical, or when most common value is needed

3.2 Measures of Spread

Range

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Limitation: Affected by extreme values (outliers).

Interquartile Range (IQR)

$$IQR = Q_3 - Q_1$$

Advantage: Not affected by extreme values.

Variance and Standard Deviation

$$\text{Variance} = s^2 = \frac{\sum(x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2$$

$$\text{Standard Deviation} = s = \sqrt{s^2}$$

Alternative Formula (often easier): $s^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2$

Example: Calculate variance for: 2, 4, 6, 8

$$\bar{x} = 5$$

x	$x - \bar{x}$	$(x-\bar{x})^2$
2	-3	9
4	-1	1
6	1	1
8	3	9
		$\sum = 20$

$$s^2 = \frac{20}{4} = 5 s = \sqrt{5} = 2.24$$

3.3 Quartiles for Grouped Data

Using Linear Interpolation:

$$Q_1 = L + \frac{(\frac{n}{4} - F)c}{f}$$

Where:

- L = lower boundary of quartile class
- F = cumulative frequency before quartile class
- f = frequency of quartile class
- c = class width

Example: Find Q_1 from:

Class	Frequency	Cumulative Frequency
0-10	8	8
10-20	15	23
20-30	12	35
30-40	5	40

$$\frac{n}{4} = 10$$

Q_1 is in 10-20 class (cumulative goes from 8 to 23)

$$Q_1 = 10 + \frac{(10 - 8) \times 10}{15} = 10 + \frac{20}{15} = 10 + 1.33 = 11.33$$

3.4 Skewness

Using Quartiles:

Type	Relationship
Positive Skew	$Q_3 - Q_2 > Q_2 - Q_1$(right tail)
Negative Skew	$Q_3 - Q_2 < Q_2 - Q_1$(left tail)
Symmetrical	$Q_3 - Q_2 = Q_2 - Q_1$

Using Mean, Median, Mode:

Type	Relationship
Positive Skew	Mean > Median > Mode
Negative Skew	Mean < Median < Mode
Symmetrical	Mean = Median = Mode

TOPIC 4: PROBABILITY

4.1 Basic Probability

$$P(A) = \frac{\text{Number of outcomes in A}}{\text{Total number of outcomes}}$$

Rules:

- $0 \leq P(A) \leq 1$
- $P(A') = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4.2 Venn Diagrams

Key Formulas:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example: If $P(A) = 0.4,\ P(B) = 0.3,\ P(A \cap B) = 0.15$, find $P(A \cup B)$

Solution: $P(A \cup B) = 0.4 + 0.3 - 0.15 = 0.55$

4.3 Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Example: In a school, 60% of students play football, 40% play basketball, and 20% play both. If a student plays football, what's the probability they also play basketball?

Solution: $P(B \mid F) = \frac{P(B \cap F)}{P(F)} = \frac{0.2}{0.6} = \frac{1}{3}$

4.4 Independent Events

Definition: A and B are independent if: $P(A \cap B) = P(A) \times P(B)$

This also means: $P(A \mid B) = P(A)$and$P(B \mid A) = P(B)$

Example: Prove that if $P(A) = 0.5$, $P(B) = 0.3$, and $P(A \cup B) = 0.65$, then A and B are independent.

Solution: $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.5 + 0.3 - 0.65 = 0.15$

$$P(A) \times P(B) = 0.5 \times 0.3 = 0.15$$

Since $P(A \cap B) = P(A) \times P(B)$, A and B are independent. ✓

4.5 Mutually Exclusive Events

Definition: A and B are mutually exclusive if: $P(A \cap B) = 0$

This means: $P(A \cup B) = P(A) + P(B)$

4.6 Tree Diagrams

Key Rules:

1. Multiply along branches
2. Add for final probabilities
3. Remember to adjust probabilities for "without replacement"

Example: A bag contains 3 red and 2 blue balls. Two balls are drawn without replacement. Find the probability of drawing one of each colour.

Solution:

```
<TEXT>
First Ball          Second Ball          Probability

Red (3/5) ----→ Red (2/4)          3/5 × 2/4 = 6/20
        |
        └─────→ Blue (2/4)         3/5 × 2/4 = 6/20
Blue (2/5) ----→ Red (3/4)         2/5 × 3/4 = 6/20
        |
```

```
        └──────→ Blue (1/4)          2/5 × 1/4 = 2/20
```

$$P(\text{one of each}) = \frac{6}{20} + \frac{6}{20} = \frac{12}{20} = \frac{3}{5}$$

TOPIC 5: CORRELATION AND REGRESSION

5.1 Scatter Diagrams

Types of Correlation:

```
<TEXT>
Positive Correlation          Negative Correlation          No Correlation
     ↗                             ↘                          → →
    ↗                             ↘
   ↗                             ↘
  ↗                             ↘
```

5.2 Product Moment Correlation Coefficient (PMCC)

Formula: $r = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}}$

Where: $S_{xy} = \sum xy - \frac{\Sigma x \Sigma y}{n}$ $S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$ $S_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$

Properties:

- $-1 \leq r \leq 1$
- $r = 1$: Perfect positive correlation
- $r = -1$: Perfect negative correlation
- $r = 0$: No linear correlation

Interpretation Guide:

r value	Interpretation
0 to 0.2	Very weak/no correlation
0.2 to 0.4	Weak correlation
0.4 to 0.6	Moderate correlation
0.6 to 0.8	Strong correlation
0.8 to 1.0	Very strong correlation

Example: Calculate r for:

x	y	xy	x^2	y^2
1	2	2	1	4

x	y	xy	x^2	y^2
2	4	8	4	16
3	5	15	9	25
4	7	28	16	49
5	8	40	25	64
∑15	26	93	55	158

$$n = 5$$

$$S_{xy} = 93 - \frac{15 \times 26}{5} = 93 - 78 = 15 S_{xx} = 55 - \frac{15^2}{5} = 55 - 45 = 10 S_{yy} = 158 - \frac{26^2}{5} = 158 - 135.2 = 22.8$$

$$r = \frac{15}{\sqrt{10 \times 22.8}} = \frac{15}{\sqrt{228}} = \frac{15}{15.1} = 0.99$$

Strong positive correlation

5.3 Regression Lines

Regression Line of y on x

$$y = a + bx$$

Where: $b = \frac{S_{xy}}{S_{xx}} a = \bar{y} - b\bar{x}$

Important Notes:

- x is the explanatory (independent) variable
- y is the response (dependent) variable
- The line always passes through $(\bar{x}, \bar{y})$

Interpretation of b:

- b is the gradient
- For every 1 unit increase in x, y changes by b units

Interpolation vs Extrapolation:

- **Interpolation** (within data range): Generally reliable
- **Extrapolation** (outside data range): Risky - may not be valid

Example: Find the regression line y on x for the previous data.

Solution: $\bar{x} = \frac{15}{5} = 3, \bar{y} = \frac{26}{5} = 5.2$

$$b = \frac{15}{10} = 1.5a = 5.2 - 1.5 \times 3 = 5.2 - 4.5 = 0.7$$

$$\therefore y = 0.7 + 1.5x$$

TOPIC 6: DISCRETE RANDOM VARIABLES

6.1 Definitions

Random Variable: A variable that takes numerical values determined by the outcome of a random experiment.

Discrete Random Variable: Can only take specific, separate values.

Probability Distribution: Lists all possible values and their probabilities.

6.2 Probability Functions

$$P(X = x) = p(x)$$

Requirements:

- $p(x) \geq 0$for all x
- $\sum p(x) = 1$

6.3 Expectation (Mean)

$$E(X) = \mu = \sum x \cdot p(x)$$

Example: A fair die is rolled. Let X be the score. Find E(X).

x	**1**	**2**	**3**	**4**	**5**	**6**
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{21}{6} = 3.5$$

6.4 Variance

$$\text{Var}(X) = E(X^2) - [E(X)]^2$$

Where: $E(X^2) = \sum x^2 \cdot p(x)$

Example: Find Var(X) for the die example.

x	x^2	P(X=x)	$x^2 \times$ P(X=x)
1	1	1/6	1/6
2	4	1/6	4/6
3	9	1/6	9/6
4	16	1/6	16/6
5	25	1/6	25/6
6	36	1/6	36/6

$$E(X^2) = \frac{1+4+9+16+25+36}{6} = \frac{91}{6} = 15.167$$

$$\text{Var}(X) = 15.167 - 3.5^2 = 15.167 - 12.25 = 2.917$$

6.5 Expectation Algebra

For any constants a and b:

$$E(aX+b) = aE(X) + b$$

$$\text{Var}(aX+b) = a^2\text{Var}(X)$$

Note: Adding a constant doesn't change variance.

6.6 Cumulative Distribution Function

$$F(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} p(x)$$

Example: Given p(x):

x	0	1	2	3
p(x)	0.1	0.3	0.4	0.2

Find F(2):

$$F(2) = P(X \leq 2) = 0.1 + 0.3 + 0.4 = 0.8$$

6.7 Discrete Uniform Distribution

Definition: A discrete random variable where each value has equal probability.

For X taking values {1, 2, 3, ..., n}:

$$E(X) = \frac{n+1}{2}$$

$$\text{Var}(X) = \frac{n^2 - 1}{12}$$

TOPIC 7: THE NORMAL DISTRIBUTION

7.1 Introduction

The Normal distribution is a continuous distribution with:

- Bell-shaped curve
- Symmetrical about the mean
- Mean = Median = Mode

Notation: $X \sim N(\mu, \sigma^2)$

Where:

- μ= mean
- σ^2= variance

7.2 Standard Normal Distribution

$$Z = \frac{X - \mu}{\sigma}$$

Where $Z \sim N(0,1)$

7.3 Using Normal Distribution Tables

The tables give $\Phi(z) = P(Z \leq z)$

Key Values to Remember:

- $\Phi(0) = 0.5$
- $\Phi(1.96) \approx 0.975$
- $\Phi(1.645) \approx 0.95$
- $\Phi(2.576) \approx 0.995$

7.4 Solving Normal Distribution Problems

Example 1: If $X \sim N(100, 15^2)$, find $P(X < 115)$.

Solution: $Z = \frac{115-100}{15} = \frac{15}{15} = 1 P(X < 115) = P(Z < 1) = \Phi(1) = 0.8413$

Example 2: If $X \sim N(100, 15^2)$, find $P(X > 85)$.

Solution: $Z = \frac{85-100}{15} = \frac{-15}{15} = -1 P(X > 85) = P(Z > -1) = 1 - \Phi(-1) = 1 - (1 - \Phi(1)) = \Phi(1) = 0.8413$

Example 3: If $X \sim N(100, 15^2)$, find $P(90 < X < 110)$.

Solution: $Z_1 = \frac{90-100}{15} = -0.667 Z_2 = \frac{110-100}{15} = 0.667$

$$P(90 < X < 110) = \Phi(0.667) - \Phi(-0.667) = 0.7475 - (1 - 0.7475) = 0.495$$

Example 4: For $X \sim N(\mu, \sigma^2)$where $P(X < 50) = 0.2$and $P(X > 80) = 0.1$, find $\mu$and σ.

Solution: From tables: $P(Z < z_1) = 0.2 \Rightarrow z_1 = -0.8416 P(Z > z_2) = 0.1 \Rightarrow \Phi(z_2) = 0.9 \Rightarrow z_2 = 1.2816$

$$\frac{50 - \mu}{\sigma} = -0.8416 (1) \frac{80 - \mu}{\sigma} = 1.2816 (2)$$

From (1): $50 - \mu = -0.8416\sigma$From (2): $80 - \mu = 1.2816\sigma$

Subtracting: $30 = 2.1232\sigma\sigma = 14.13$

Substituting: $50 - \mu = -0.8416 \times 14.13\mu = 50 + 11.89 = 61.89$

EXAM TIPS FOR S1

Common Mistakes to Avoid

1. **Normal distribution:** Always standardise using $Z = \frac{X-\mu}{\sigma}$
2. **Probability:** Make sure events are independent before multiplying
3. **Variance formula:** Remember it's $E(X^2) - [E(X)]^2$, not $E(X^2 - X)^2$
4. **Interpolation/extrapolation:** Don't extrapolate regression lines beyond data range
5. **Box plots:** Always show a scale and label axes

Calculator Tips

1. Use statistical mode for standard deviation and regression
2. Store intermediate values to avoid rounding errors
3. Know how to use your calculator's normal distribution functions

Answer Presentation

1. Give probabilities to 4 decimal places
2. Give means to appropriate accuracy (often 3 significant figures)
3. State hypotheses clearly in hypothesis testing
4. Draw diagrams where helpful (especially for probability and normal distribution)

Time Management

- ~2 minutes per mark
- Don't spend too long on one question
- If stuck, move on and return later

PRACTICE PROBLEMS

Probability

1. A bag contains 4 red and 6 blue balls. Two balls are drawn without replacement. Find the probability that both are red.
2. If $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cup B) = 0.7$, find $P(A \cap B)$and determine if A and B are independent.

Statistics

3. Find the mean, median, mode, and standard deviation for: 12, 15, 18, 20, 22, 25, 28
4. Draw a box plot for: 8, 12, 15, 17, 19, 22, 25, 30, 35

Normal Distribution

5. If $X \sim N(50, 8^2)$, find: a) $P(X < 45)$b) $P(X > 60)$c) $P(40 < X < 55)$

Correlation and Regression

6. Calculate the PMCC for:

x	**1**	**2**	**3**	**4**	**5**
y	3	5	6	8	9

7. Find the regression line y on x for the data in question 6.

Discrete Random Variables

8. A random variable X has probability distribution:

x	**0**	**1**	**2**	**3**
P(X=x)	0.1	0.3	0.4	0.2

Find: a) $E(X)$b) Var(X)c) $F(2)$

ANSWERS TO PRACTICE PROBLEMS

1. Probability

$$P(\text{both red}) = \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = \frac{2}{15}$$

2. Independence Check

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

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

3. Statistics

Mean = $\frac{140}{7} = 20$Median = 20 (middle value) Mode = no mode Standard deviation = 6.06

4. Box Plot Values

Minimum = 8, Q_1 = 15, Median = 19, Q_3 = 25, Maximum = 35

5. Normal Distribution

a) $Z = \frac{45-50}{8} = -0.625,\ P = 0.2659$b) $Z = \frac{60-50}{8} = 1.25,\ P = 0.1056$c) $Z_1 = -1.25,\ Z_2 = 0.625,\ P = 0.6293$

6. PMCC

$r = 0.986$(strong positive correlation)

7. Regression Line

$$y = 1.9 + 1.5x$$

8. Discrete Random Variables

a) $E(X) = 1.7$b) $\text{Var}(X) = 0.81$c) $F(2) = 0.8$