Statistics and Probability

Matthew Williams • Add Math • May 16, 2026

Statistics and Probability

Section 4 contributes 5 items to Paper 01 and one 20-mark question to Paper 02. The Paper 02 question typically combines statistical calculation with interpretation and a probability problem using diagrams.

Types of Data

Qualitative (categorical) data describes attributes without numerical value: eye colour, nationality, type of vehicle.

Quantitative data is numerical and divides into two types:

- **Discrete:** takes separate countable values (number of siblings, goals scored).
- **Continuous:** takes any value within an interval (height, temperature, time).

Measures of Central Tendency

Measure	Definition	Best used when
Mean	Sum of all values divided by the count	Data is symmetric with no extreme outliers
Median	Middle value when data is ordered	Data is skewed or has outliers
Mode	Most frequent value	Identifying the most common category

$$\bar{x} = \frac{\sum x}{n} \quad \text{(ungrouped)} \qquad \bar{x} = \frac{\sum fx}{\sum f} \quad \text{(grouped/frequency)}$$

Measures of Spread

The simplest measure of spread is the **range**: $\text{range} = \text{maximum} - \text{minimum}$. It is easy to compute but sensitive to a single extreme value.

Quartiles and Interquartile Range

Ordered data is divided into four equal parts by the quartiles $Q1$, $Q2$ (median), and Q_3.

$$Q_1 = \text{median of lower half}, \quad Q_2 = \text{median}, \quad Q_3 = \text{median of upper half}$$

$$\text{IQR} = Q_3 - Q_1 \qquad \text{Semi-IQR} = \frac{Q_3 - Q_1}{2}$$

The IQR measures the spread of the **middle 50%** of the data. Unlike the range, it is not distorted by a single extreme value.

> 📋 **Example**
> Data (already ordered): $12, 15, 18, 20, 22, 25, 27, 30, 32$.
>
> Median (Q_2): 5th value $= 22$.
>
> Lower half: $12, 15, 18, 20$. $Q_1 = \frac{15 + 18}{2} = 16.5$
>
> Upper half: $25, 27, 30, 32$. $Q_3 = \frac{27 + 30}{2} = 28.5$
>
> $\text{IQR} = 28.5 - 16.5 = 12$

Percentiles

The pth percentile is the value below which $p\%$ of the data lies. Quartiles are specific percentiles: $Q1 = P25$, $Q2 = P50$, $Q3 = P75$.

Variance and Standard Deviation

Variance measures average squared distance from the mean. **Standard deviation** is its square root, restoring the original units.

$$S^2 = \frac{\sum(x - \bar{x})^2}{n} \qquad S = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} \qquad \text{(ungrouped)}$$

$$S^2 = \frac{\sum f(x - \bar{x})^2}{\sum f} \qquad \text{(grouped)}$$

A **small** standard deviation means values cluster closely around the mean (consistent). A **large** standard deviation means values are widely spread (variable).

> 📝 **Exam Tip**
> Two datasets can have the same mean but very different standard deviations. When comparing distributions, state both the measure of centre and the measure of spread. "Group A has a higher mean but Group B is more consistent because its standard deviation is smaller."

Stem-and-Leaf Diagrams

A stem-and-leaf diagram keeps the original data values while showing the distribution shape.

Constructing one:

- 1. Split each value: the stem is all digits except the last; the leaf is the last digit.
- 2. Write stems in a column and leaves in ascending order beside each stem.
- 3. Include a key.

> 📋 **Example**
> Data: $17, 19, 22, 31, 33, 38, 44, 47, 49$
>
> [CodeBlock:0]

A **back-to-back** stem-and-leaf diagram places two datasets side by side sharing a common stem, allowing direct visual comparison.

Advantages of stem-and-leaf	Disadvantages

Original values are preserved	Impractical for large datasets
Shape of distribution is visible	Difficult to compare non-overlapping ranges

Box-and-Whisker Plots

A box plot summarises a dataset using five values: minimum, $Q1$, median, $Q3$, maximum.

[Code: plaintext]

```
Min    Q•    Q,    Qf    Max
|------[======|======]------|
```

The **box**spans from $Q1$ to $Q3$ and contains the middle 50% of the data. The **whiskers** extend to the minimum and maximum.

Reading Skewness from a Box Plot

Pattern	Skew
[Math: Q3 - Q2 = Q2 - Q1] and whiskers roughly equal	Symmetric
[Math: Q3 - Q2 > Q2 - Q1] or right whisker longer	Positive skew (tail to right)
[Math: Q3 - Q2 < Q2 - Q1] or left whisker longer	Negative skew (tail to left)

For a positively skewed distribution: Mode $<$ Median $<$ Mean.

For a negatively skewed distribution: Mean $<$ Median $<$ Mode.

Probability Theory

An **experiment**produces **outcomes**. The set of all possible outcomes is the **sample space** S. An **event**is a subset of the sample space.

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes in } S}$$

This is **classical probability**, valid when all outcomes are equally likely.

Relative frequency gives an experimental estimate of probability when theoretical equally-likely outcomes cannot be assumed:

$$P(A) \approx \frac{\text{number of times } A \text{ occurred}}{\text{total number of trials}}$$

As the number of trials increases, the relative frequency approaches the true probability.

Basic Laws

- $0 \leq P(A) \leq 1$ for any event A.
- $\sum P(\text{all outcomes}) = 1$.
- $P(A') = 1 - P(A)$, where A' is the complement of A (the event that A does not occur).

The Addition Rule

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

For **mutually exclusive** events (A and B cannot both occur): $P(A \cap B) = 0$, so:

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

📋 **Example**

A card is drawn from a standard deck. $P(K) = 4/52$ and $P(A) = 4/52$ where K = King and A = Ace. Since these are mutually exclusive:

$$P(K \cup A) = \frac{4}{52} + \frac{4}{52} = \frac{8}{52} = \frac{2}{13}$$

Conditional Probability

The conditional probability of A given B has occurred:

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

Rearranging: $P(A \cap B) = P(A|B) \cdot P(B)$.

Independent Events

Events A and B are **independent** if knowing that B occurred gives no information about A:

$$P(A|B) = P(A) \qquad \textbf{equivalently} \qquad P(A \cap B) = P(A) \cdot P(B)$$

> 📝 **Exam Tip**
> Do not confuse "mutually exclusive" with "independent." Mutually exclusive events ($P(A \cap B) = 0$) cannot both occur. Independent events can both occur, but neither influences the other's probability. Two events with non-zero probabilities cannot be both mutually exclusive and independent.

Probability Diagrams

Possibility Space Diagrams

A possibility space (sample space) diagram lists all outcomes in a grid. Useful for two-stage experiments like rolling two dice.

> 📋 **Example**
> Two fair dice are rolled. The sample space has $6 \times 6 = 36$ equally likely outcomes.
>
> $P(\text{sum} = 7) = \frac{6}{36} = \frac{1}{6}$ (the six pairs that sum to 7: $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$).

Tree Diagrams

Tree diagrams show sequential outcomes. Branches show each possible outcome at each stage; multiply along branches for intersection probabilities.

The syllabus restricts tree diagrams to **two initial branches**.

📋 **Example**

A bag contains 3 red and 5 blue balls. A ball is drawn, its colour noted, then a second ball is drawn without replacement.

P(both red) $= \frac{3}{8} \times \frac{2}{7} = \frac{6}{56} = \frac{3}{28}$

P(one of each) $= \frac{3}{8} \times \frac{5}{7} + \frac{5}{8} \times \frac{3}{7} = \frac{15}{56} + \frac{15}{56} = \frac{30}{56} = \frac{15}{28}$

Venn Diagrams

Venn diagrams with two sets partition the sample space into four regions: A only, B only, $A \cap B$, and neither. The sum of all regions must equal $P(S) = 1$ (or $n(S)$ for counts).

The syllabus restricts Venn diagrams to **two sets**.

📋 **Example**

In a class of 30 students, 18 study Physics, 15 study Chemistry, and 8 study both.

P(Physics only) $= \frac{18 - 8}{30} = \frac{10}{30}$

P(at least one) $= \frac{18 + 15 - 8}{30} = \frac{25}{30} = \frac{5}{6}$

P(neither) $= \frac{30 - 25}{30} = \frac{5}{30} = \frac{1}{6}$