Statistics and probability

Pearson Edexcel · International GCSE · Mathematics A

4MA1 topic 6 (specification Issue 2, November 2017). Foundation 6.1–6.3 covers presenting data, averages with grouped estimates, and the full probability toolkit (scale, sample spaces, Venn diagrams, complements); Higher adds histograms with unequal classes, cumulative frequency work with median and interquartile range, and tree diagrams for independent and conditional events. On the June 2024 papers this strand supplied an unequal-class histogram (1HR Q16), a cumulative-frequency graph with quartile reads (2HR Q11) and the algebra-heavy without-replacement question (1HR Q21).

Must-know for topic 6:

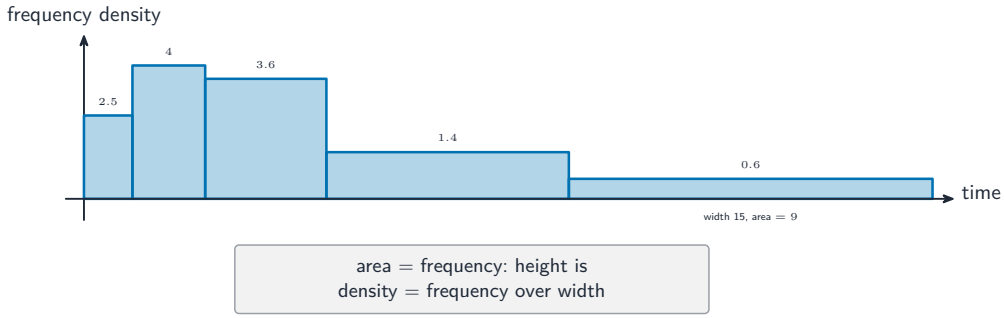
Family	Formulae
averages	mean $= \dfrac{\sum x}{n}$; grouped mean $= \dfrac{\sum fm}{\sum f}$ (m = midpoints)
histogram	frequency density $= \dfrac{\text{frequency}}{\text{class width}}$ – bar AREA is the frequency
cumulative frequency	plot at the UPPER class boundary; median at $\frac{n}{2}$, quartiles at $\frac{n}{4}$ and $\frac{3n}{4}$
probability	$P(A') = 1 - P(A)$; $P(A \cup B) = P(A) + P(B) - P(A \cap B)$; independent: multiply

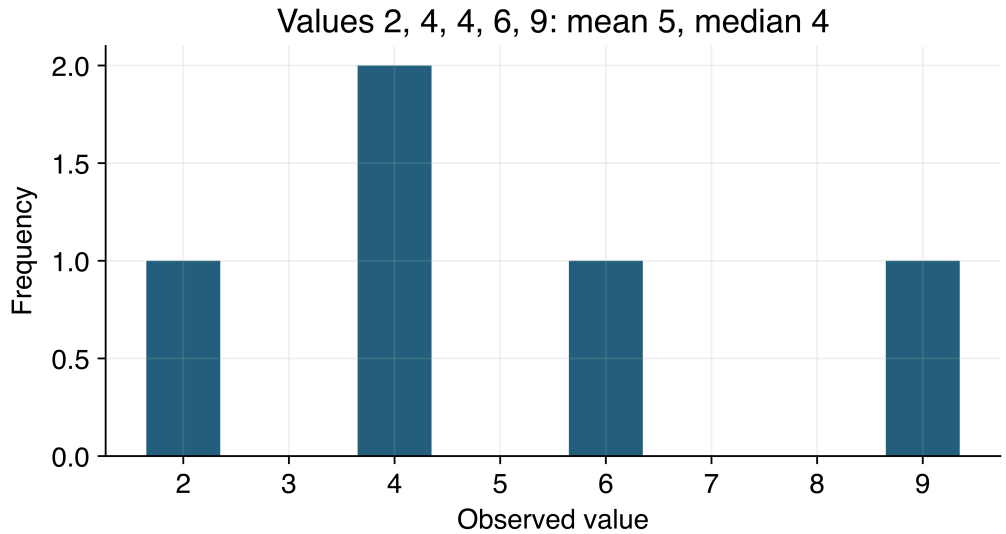
Collecting and presenting data

Tabulate before drawing: tally into a frequency table, keep class boundaries explicit, then choose the diagram – bar chart (discrete categories), pie chart (proportions), pictogram, or a frequency polygon for grouped data. Interpretation runs both ways: read values off a given chart and compare distributions (two groups on the same axes; a truncated axis that misleads is also fair comment).

Higher draws **histograms** with **unequal class widths**: the bar's height is the **frequency density** $= \dfrac{\text{frequency}}{\text{class width}}$, so the AREA represents the frequency. Equal-width classes are the special case where height alone works.

Worked example (1HR June 2024 Q16, adapted). Classes of widths 2, 3, 5, 10, 15 hold frequencies 5, 12, 18, 14, 9. Densities: $\frac{5}{2} = 2.5$, $\frac{12}{3} = 4$, $\frac{18}{5} = 3.6$, $\frac{14}{10} = 1.4$, $\frac{9}{15} = 0.6$ – the paper's marking accepts bars of correct width whose heights are in this ratio. Check yourself: the last bar looks tall only because its width is 15; its area is $0.6 \times 15 = 9$. ✓





Original native-lesson illustration; labels belong to its worked example.

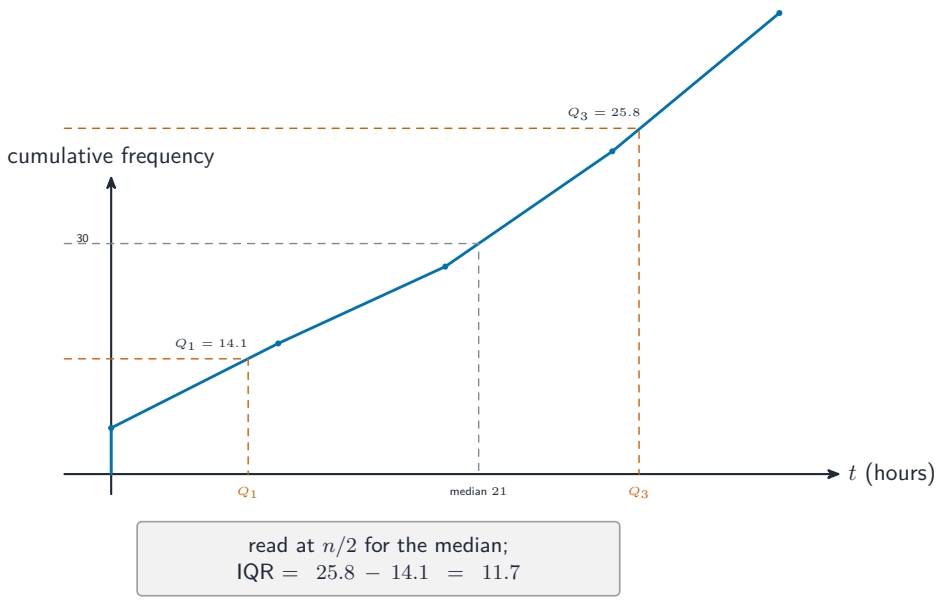
Averages and spread

The **mean** sums and divides; the **median** orders and takes the middle; the **mode** is the most frequent. For 3, 7, 8, 10, 12: mean 8, median 8, range $12 - 3 = 9$. For grouped data, estimate the mean with midpoints: multiply each midpoint by its frequency, sum, divide by the total; state the **modal class** (the class with the highest frequency density when widths differ).

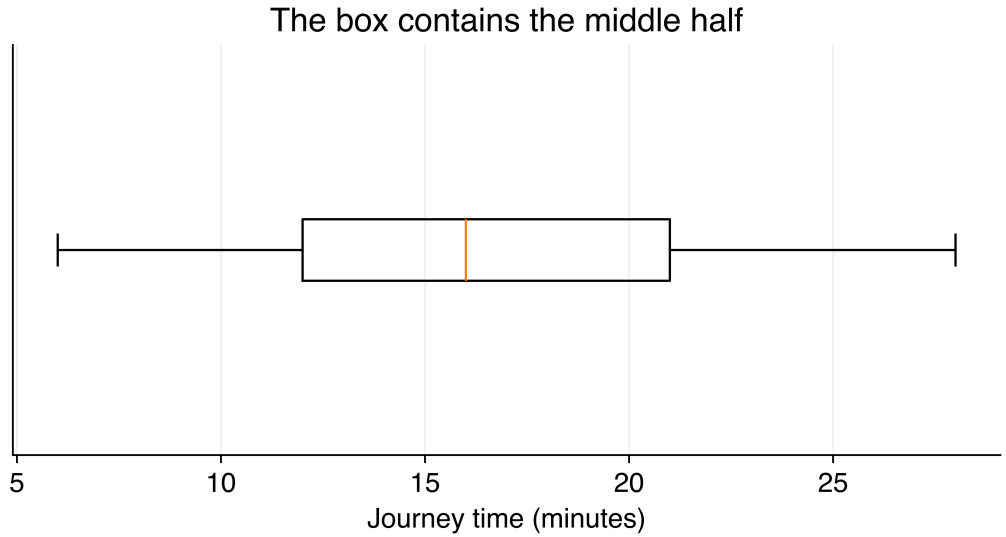
Higher quantifies **spread** with the **quartiles**: order the data, take the median, then the medians of the lower and upper halves give Q_1 and Q_3; the **interquartile range** is $Q_3 - Q_1$ (resistant to outliers, unlike the range). For 4, 6, 7, 9, 10, 12, 14, 15: the median is 9.5, the lower half 4, 6, 7, 9 gives $Q_1 = 6.5$, the upper half 10, 12, 14, 15 gives $Q_3 = 13$, so the IQR is $13 - 6.5 = 6.5$.

Cumulative frequency diagrams

A **cumulative frequency** table adds frequencies as classes pass; plot each total against its UPPER class boundary and join with a smooth increasing curve – the **ogive**. Reads: the median is the value at $\frac{n}{2}$ on the curve; the quartiles at $\frac{n}{4}$ and $\frac{3n}{4}$; the IQR is their difference; percentiles work the same way, and "how many above X" reads the total at X and subtracts from n.



Worked example (2HR June 2024 Q11 family, adapted). 60 workers' homework times (hours per week): cumulative frequencies 6, 17, 27, 42, 60 at upper boundaries 10, 15, 20, 25, 30. The median sits at the 30th worker: between 20 (27 workers) and 25 (42), so $20 + 5 \times \frac{30-27}{42-27} = 21$ hours. The quartiles: the 15th worker gives $Q_1 = 10 + 5 \times \frac{15-6}{17-6} = 14.1$; the 45th gives $Q_3 = 25 + 5 \times \frac{45-42}{60-42} = 25.8$; IQR ≈ 11.7. Draw the reading lines on the graph – the marks are for the reads, not just the answers.

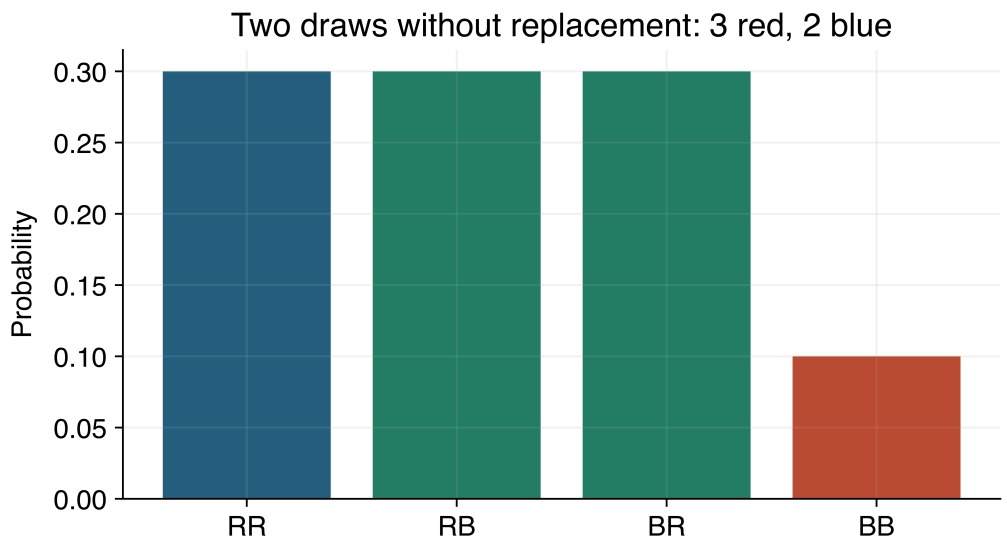


Original native-lesson illustration; labels belong to its worked example.

Probability foundations

Probability measures likelihood on the scale 0 (impossible) to 1 (certain). A theoretical model counts equally likely outcomes; an experiment estimates from frequency: $P(\text{event}) \approx \frac{\text{times it happened}}{\text{trials}}$ (estimates from previously collected data included). The **sample space** lists all outcomes – for two successive events, a grid or systematic list; an event's probability counts the favourable cells. The **complement** rule: $P(A') = 1 - P(A)$ – often the fast route ("at least one" questions).

Venn diagrams : regions for each set and the overlap; probabilities fill the regions and sum to 1, with $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. Example: 30 students, 18 like cricket (C), 15 tennis (T), 7 both: $P(C \cup T) = \frac{18+15-7}{30} = \frac{26}{30}$, and $P(\text{neither}) = 1 - \frac{26}{30} = \frac{2}{15}$.



Original native-lesson illustration; labels belong to its worked example.

Tree diagrams and combined events

Tree diagrams multiply along branches and add between outcomes. WITH replacement, branches repeat – the events are **independent** : $P(A \cap B) = P(A) \times P(B)$. WITHOUT replacement, the second branch's denominator drops by one and its numerator by the first outcome: the draws are dependent, and the working shows the changing fractions.

Worked example (1HR June 2024 Q21 family, adapted). A bag holds 5 red, 6 blue and 9 green sweets; two are taken without replacement. $P(\text{one red and one blue, in either order}) = \frac{5}{20} \times \frac{6}{19} + \frac{6}{20} \times \frac{5}{19} = \frac{60}{380} = \frac{3}{19}$. The paper's version runs the same machinery with an unknown x and solves the resulting quadratic (its verified results: $x = 11$, final probability $\frac{7}{75}$): set the two-route expression equal to the given probability, expand, solve, and reject the root the context forbids.

Conditional probability shrinks the world: $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)}$ – on a tree, the second branch given the first happened. With the bag above, $P(\text{blue second} \mid \text{red first}) = \frac{6}{19}$ read straight off the branch.

Checklist

- Tabulate and present data; interpret charts and compare distributions (F 6.1 A–C).
- Draw histograms with unequal widths via frequency density; area is the frequency (H 6.1 A; June 2024 1HR Q16).
- Mean, median, mode, range; grouped mean via midpoints; modal class (F 6.2 B–D).
- Quartiles and IQR for discrete data by position (H 6.2 B–C).
- Construct cumulative frequency curves from tables; read medians, quartiles, IQR and percentiles with drawn reading lines (H 6.1 B–C, H 6.2 A, D; June 2024 2HR Q11).
- Probability scale; theoretical vs experimental estimates; sample spaces and systematic listing for two events; Venn diagrams with the addition rule; complements (F 6.3 A–H).
- Tree diagrams: multiply along, add between; with replacement (independent) vs without; conditional probability read from branches; algebraic setups ending in quadratics (H 6.3 A–D; June 2024 1HR Q21 family).

Boundary notes: histograms use frequency density (not frequency) whenever widths differ; quartile conventions here are positional (medians of the lower and upper halves); expected value and formal set notation belong to later study. Foundation candidates meet 6.1–6.3 Foundation rows only. Worked examples are adapted from the reviewed June 2024 papers or original; sheet questions are original unless labelled.