

Representation of data

A-Level Maths 9709 · Topic 5.1 revision notes

Everything from the 3 lessons of this topic on a few pages: the key idea of each lesson, the diagram to remember, the points to learn and the traps examiners set. Read it, cover it, test yourself.

1. Stem-and-leaf, box plots, histograms and cumulative frequency.
2. Measures of central tendency and variation
3. Variance and standard deviation, including coded data

thelucidstem.com

1 Stem-and-leaf, box plots, histograms and cumulative frequency

Syllabus 5.1

In a histogram the AREA of a bar is the frequency, so the height is the frequency density: $\text{frequency density} = \frac{\text{frequency}}{\text{class width}}$. On a cumulative frequency graph, read across from a position (such as $\frac{n}{2}$ for the median) and down to the value.

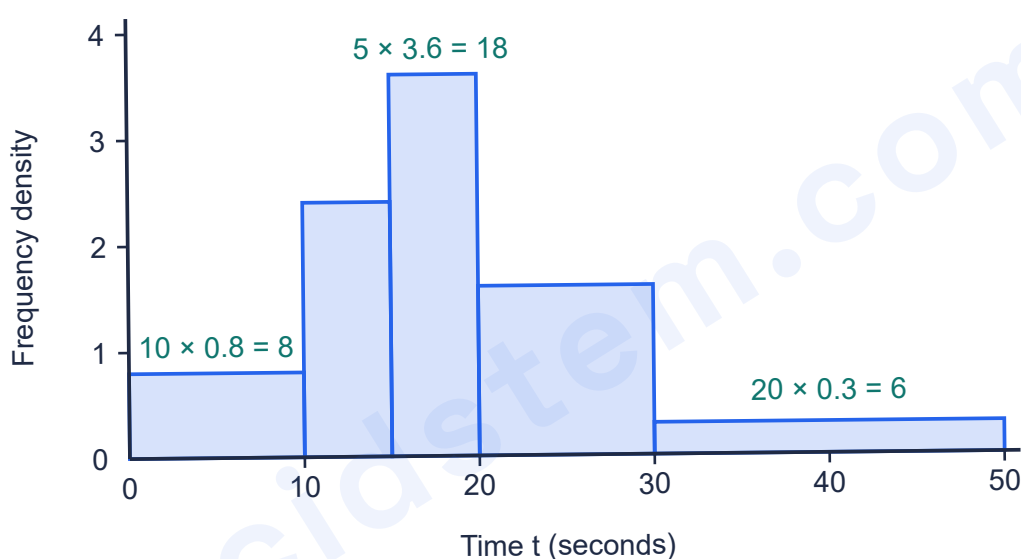


Fig. 1: a histogram with unequal class widths. Each bar's area (width $\times$ frequency density) is its frequency: 8, 12, 18, 16 and 6, total 60

Worked example: Frequency densities

The masses, to the nearest kilogram, of 57 sacks are grouped as 40-44 (15 sacks), 45-54 (30 sacks) and 55-59 (12 sacks). Find the frequency density for each class.

1. Class boundaries: 39.5 to 44.5, 44.5 to 54.5 and 54.5 to 59.5.
2. Class widths: 5, 10 and 5.
3. Frequency densities: $15 \div 5 = 3$, $30 \div 10 = 3$, $12 \div 5 = 2.4$.

Answer: 3, 3 and 2.4 sacks per kg

Boundaries first, then density or position.

- ✓ **Boundaries:** for rounded data, extend each class by half a unit each way before finding widths.
- ✓ **Histograms:** height = frequency $\div$ width; frequency = width $\times$ height.
- ✓ **Stem-and-leaf:** order the data, give a key for both sides, read left leaves backwards.
- ✓ **Cumulative frequency:** plot at upper boundaries; read at $\frac{n}{4}$, $\frac{n}{2}$, $\frac{3n}{4}$ or $\frac{k}{100}n$.
- ✓ **Compare:** one comment on average, one on spread, both in context.

Watch out!

Plotting frequency (not frequency density) on the vertical axis when the class widths differ. It makes wide classes look far too important and scores no marks for the heights.

Forgetting to reverse the left-hand leaves. In the row 8 6 5 2 0 | 3, the smallest Class A value is 30 (the leaf next to the stem), not 38.

Plotting cumulative frequencies at the class midpoints. A cumulative frequency counts everything UP TO the end of a class, so it belongs at the upper boundary.

2 Measures of central tendency and variation

Syllabus 5.1

The mean $\bar{x} = \frac{\sum fx}{\sum f}$ uses every value, so one extreme value can drag it. The median and interquartile range use only positions, so they resist outliers and skew. Quote a centre AND a spread, and compare them in context.

Seven salaries: 18, 20, 21, 22, 24, 25, 95

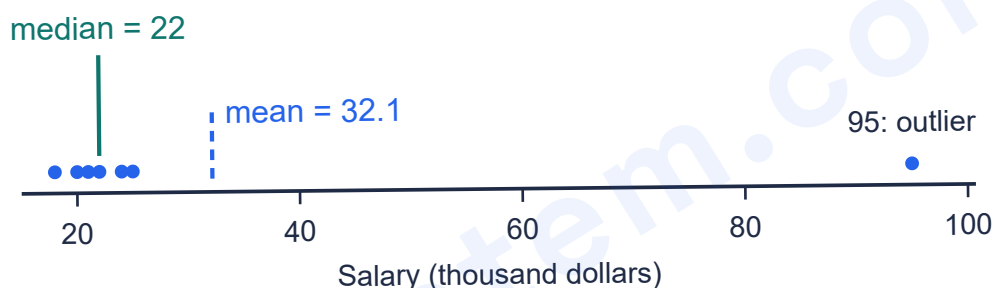


Fig. 1: one extreme salary drags the mean to 32.1, above six of the seven values. The median, 22, still sits in the middle of the typical salaries

Worked example: From a frequency table

The numbers of goals x scored in 25 matches are: $x = 0$: 5 matches, 1: 8, 2: 6, 3: 4, 4: 2. Find the mean, median and mode.

1. $\sum fx = 0 + 8 + 12 + 12 + 8 = 40$, so $\bar{x} = \frac{40}{25} = 1.6$.
2. The median is the 13th value. Cumulative frequencies: 5, 13, 19, 23, 25. The 13th value is 1.
3. The highest frequency is 8, at $x = 1$.

Answer: Mean 1.6, median 1, mode 1

Totals for means, positions for medians.

- ✓ **Mean:** $\bar{x} = \frac{\sum fx}{\sum f}$; for grouped data use midpoints.
- ✓ **Missing or combined values:** work with totals, $\sum x = n\bar{x}$.
- ✓ **Median and quartiles:** find the position, then count along the cumulative frequencies.
- ✓ **Choose:** outliers or skew means median and IQR; otherwise mean and standard deviation.
- ✓ **Compare:** one average and one spread statement, in context.

Watch out!

Averaging the two means.

$\frac{1}{2}(64 + 59) = 61.5$ is wrong because the groups are different sizes; always go back to the totals $\sum x$.

Using the upper class boundary (or the class width) instead of the midpoint in $\sum fx$. For $30 \leq t < 50$ the midpoint is 40, not 50 and not 20.

Quoting a quartile POSITION instead of a value: for 11 values Q_3 is the 9th value (26), not 9.

Saying 'Ana's data are higher' without context. In this question a lower time is better, so say 'faster', and refer to times, not just numbers.

3 Variance and standard deviation, including coded data

Syllabus 5.1

$$\text{Variance} = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2 \text{ and standard deviation} = \sqrt{\text{variance}}.$$

Coding by subtracting a shifts the mean by a but leaves the standard deviation unchanged.

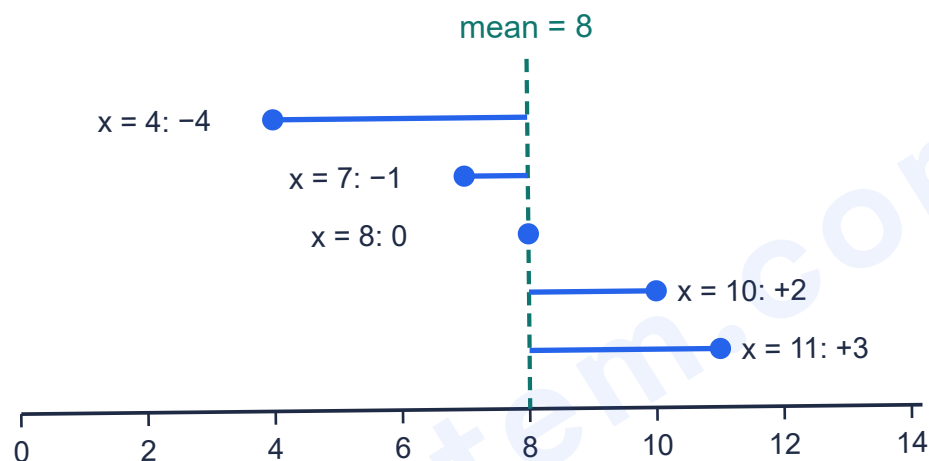


Fig. 1: deviations from the mean for 4, 7, 8, 10, 11. Squared and averaged: $(16 + 1 + 0 + 4 + 9) \div 5 = 6$, so the standard deviation is $\sqrt{6} \approx 2.45$

Worked example: From raw data, both ways

Find the standard deviation of 4, 7, 8, 10, 11.

1. $\bar{x} = \frac{40}{5} = 8.$

2. Definition: deviations $-4, -1, 0, 2, 3$; squares sum to $16 + 1 + 0 + 4 + 9 = 30$; variance $= 30 \div 5 = 6.$

3. Quick formula: $\sum x^2 = 16 + 49 + 64 + 100 + 121 = 350$; variance $= \frac{350}{5} - 8^2 = 70 - 64 = 6.$

4. Standard deviation $= \sqrt{6}.$

Answer: $\sqrt{6} = 2.45$ (3 s.f.)

Get to n , $\sum x$ and $\sum x^2$.

- ✓ **Mean:** $\bar{x} = \frac{\sum x}{n}$ (or $\frac{\sum fx}{\sum f}$).
- ✓ **Variance:** $\frac{\sum x^2}{n} - \bar{x}^2$; standard deviation is its square root.
- ✓ **Coded:** $\text{mean} = a + \frac{\sum(x - a)}{n}$; variance is the same as for the coded values.
- ✓ **Reverse:** $\sum x = n\bar{x}$ and $\sum x^2 = n(\sigma^2 + \bar{x}^2)$.
- ✓ **Combine:** add the totals, then use the formulas once.

Watch out!

Writing $\frac{\sum x^2}{n} - \bar{x}$ or $\frac{(\sum x)^2}{n}$.

It is the MEAN that is squared and subtracted: $\frac{\sum x^2}{n} - \bar{x}^2$.

Using $(\sum fx)^2$ or $\sum (fx)^2$ in place of $\sum fx^2$. Square the value x first, then multiply by its frequency.

Thinking $\sum x^2 = \sum (x - a)^2 + na^2$. The cross term $-2a \sum x$ matters; expand the bracket properly or work via the variance.

The combined standard deviation (3.95) is bigger than both separate ones (2 and 3), because the two means are far apart.

Averaging the standard deviations misses this completely.