Cambridge AS & A Level · Mathematics (9709) · Notes

Representation of data AS

Read online at pastopic.com/cambridge/mathematics-9709/notes/representation-of-data

Last checked 28 September 2026

What you need to know

By the end of this topic you should be able to:

1. **Choose a sensible way to display raw data**, and say what is good or bad about a particular kind of diagram (for example, what a stem-and-leaf diagram shows that a box-and-whisker plot does not).
2. **Draw and read four kinds of diagram**: stem-and-leaf diagrams (including back-to-back ones that show two sets of data), box-and-whisker plots, histograms and cumulative frequency graphs.
3. **Know and use the measures of average** (mean, median, mode) **and of spread** (range, interquartile range, standard deviation), and use them to compare two sets of data.
4. **Read estimates from a cumulative frequency graph**: the median, the quartiles, any percentile, and how many values lie above a value, below it, or between two values.
5. **Calculate the mean and standard deviation** of raw data or grouped data, either from the values themselves or from totals such as $\sum x$ and $\sum x^2$, or coded totals $\sum(x-a)$ and $\sum(x-a)^2$. Use these totals to solve problems, including ones that put two sets of data together.

Understand it

Raw data and grouped data

Raw data is a list of the actual values, such as the masses of 11 apples. With a small set you can show every value. With a large set (hundreds of values) the data is usually **grouped** into classes, and you only know how many values are in each class, not the values themselves. Anything worked out from grouped data (a mean, a median) is therefore an **estimate**.

Measurements such as time, length and mass are **continuous**: they can take any value in a range. Counts, such as the number of goals, are **discrete**.

Stem-and-leaf diagrams

A stem-and-leaf diagram splits each value into a **stem** (the leading digits) and a **leaf** (the last digit). The leaves are written **in order**, lined up in columns, so the diagram is also a sideways bar chart. A **back-to-back** diagram shares one stem between two data sets: one set's leaves go to the left, with the smallest leaf next to the stem, and the other set's leaves go to the right.

North	g	South
	11	8
4	**12**	7 9
8 5 1	**13**	3 6 6
7 6 2	**14**	0 4 9
8 3 1	**15**	
6	**16**	1
	17	2

Key: 6 | 14 | 4 means 146 g for North and 144 g for South

Every diagram needs a **key** that shows how to read it, naming both sets and giving the units. Choose a key whose two leaves are different, like $6 \mid 14 \mid 4$, so it is clear which side is which.

Median and quartiles of raw data

Put the values in order. With n values:

- the **median** Q_2 is the middle value, the $\frac{n+1}{2}$th (for even n, halfway between the two middle values);
- the **lower quartile** Q_1 is the median of the lower half, and the **upper quartile** Q_3 is the median of the upper half. When n is odd, leave the median itself out of both halves.

North's 11 masses in order are 124, 131, 135, 138, 142, **146**, 147, 151, 153, 158, 166. The median is the 6th, 146 g. The lower half is 124, 131, 135, 138, 142, so $Q_1 = 135$; the upper half gives $Q_3 = 153$.

A quick way to find positions: for $n = 11$ the quartiles are the 3rd and 9th values; for $n = 15$ the 4th, 8th and 12th; for $n = 19$ the 5th, 10th and 15th; for $n = 27$ the 7th, 14th and 21st. On the left of a back-to-back diagram, count **outwards from the stem**, starting at the top.

Measures of average and spread

Measure	What it is	Good for
Mean	total ÷ number of values	uses every value; but pulled towards extreme values
Median	middle value	not affected by extreme values
Mode	most common value	data with a clear most common value
Range	largest – smallest	quick, but set by the two most extreme values
Interquartile range (IQR)	$Q_3 - Q_1$	spread of the middle half; ignores extreme values
Standard deviation	typical distance of the values from the mean	uses every value; pairs with the mean

For North, the range is $166 - 124 = 42$ g and the IQR is $153 - 135 = 18$ g. South's values give $Q_1 = 129$, median 136, $Q_3 = 149$, IQR 20 g and range 54 g. South's mode is 136 g, the only value that appears twice.

Which average? If the data contains an **extreme value** (an outlier) or is **skewed** (bunched at one end with a long tail at the other), the mean is pulled towards the tail, so the **median** represents the data better. If the data is roughly symmetrical, the mean is fine and uses every value.

Mean and standard deviation

The **mean** is $\bar{x} = \dfrac{\sum x}{n}$. The **variance** is the mean of the squared distances from the mean, and the **standard deviation** is its square root:

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

The second form, "mean of the squares minus the square of the mean", is the one you use in practice. For North, $\sum x = 1591$ and $\sum x^2 = 231\,625$, so $\bar{x} = 144.6$ g and the standard deviation is $\sqrt{\frac{231\,625}{11} - 144.636^2} = 11.7$ g.

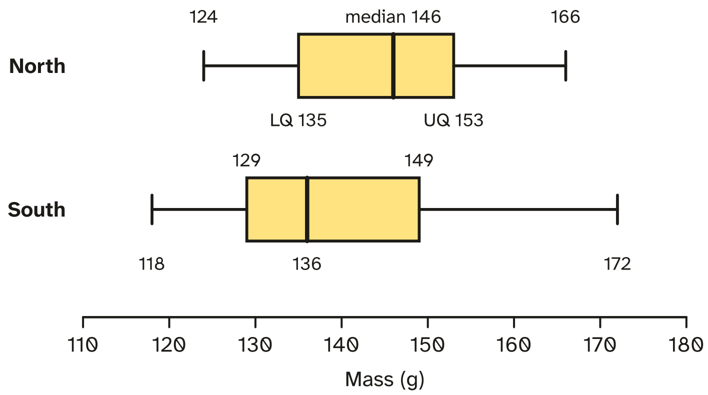
Coded data. If you are given $\sum(x-a)$ and $\sum(x-a)^2$ for some number a, work with $y = x - a$. Subtracting a from every value moves the mean down by a but does **not** change the spread:

$$\bar{x} = a + \frac{\sum(x-a)}{n}, \qquad \text{sd of } x = \sqrt{\frac{\sum(x-a)^2}{n} - \left(\frac{\sum(x-a)}{n}\right)^2}.$$

Combining two sets. Means and standard deviations cannot simply be added or averaged. Instead, get back to totals: $\sum x = n\bar{x}$ and $\sum x^2 = n(\sigma^2 + \bar{x}^2)$. Add the n's, the $\sum x$'s and the $\sum x^2$'s of the two sets, then use the formulas on the combined totals.

Box-and-whisker plots

A box-and-whisker plot shows five values: the lowest value, Q_1, the median, Q_3 and the highest value. The box runs from Q_1 to Q_3 with a line at the median, and the whiskers run out to the lowest and highest values. Two plots drawn on **one** scale are easy to compare.



North's median is higher, so North's apples are heavier on average. South's box and whiskers are longer, so South's masses are more spread out (less consistent).

Grouped data: boundaries, widths and midpoints

For a class of continuous data, find its **class boundaries** first:

- "$20 \leqslant t < 25$" has boundaries 20 and 25;
- "21–25 minutes, **to the nearest minute**" has boundaries 20.5 and 25.5, because a time of 20.6 rounds to 21;
- an age class "21–25 years" (age in whole years) runs from 21 up to 26.

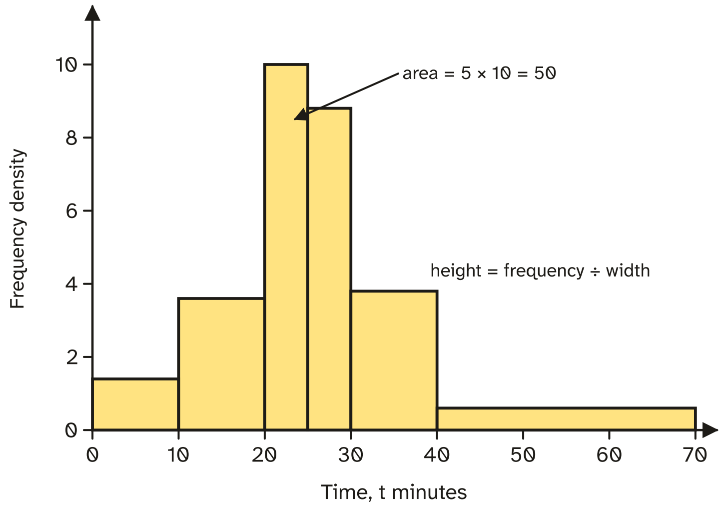
The **class width** is upper boundary minus lower boundary, and the **midpoint** is halfway between the boundaries. The midpoint stands in for every value in the class when you estimate a mean.

Histograms

In a histogram the **area** of each bar stands for the frequency, so bars can have different widths. The height of each bar is the **frequency density**:

$$\text{frequency density} = \frac{\text{frequency}}{\text{class width}}.$$

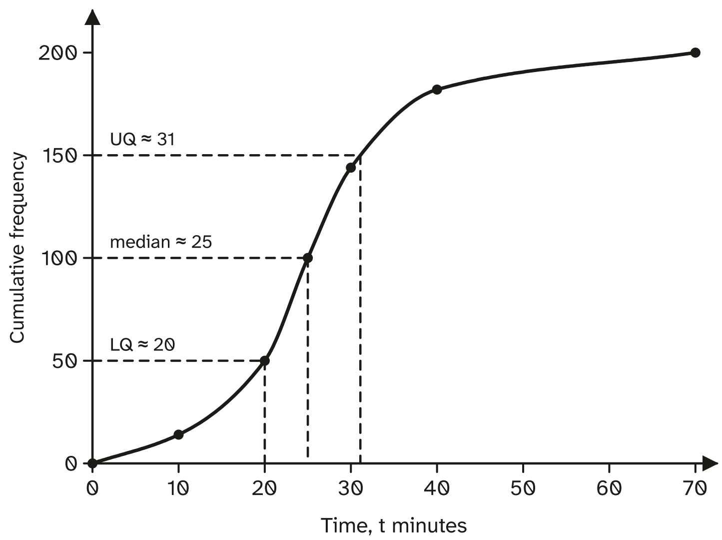
The histogram below shows the times of 200 people: $0 \leqslant t < 10$: 14 people, $10 \leqslant t < 20$: 36, $20 \leqslant t < 25$: 50, $25 \leqslant t < 30$: 44, $30 \leqslant t < 40$: 38 and $40 \leqslant t < 70$: 18. The class $40 \leqslant t < 70$ is three times as wide as $30 \leqslant t < 40$, so its 18 people make a bar only 0.6 high. A bar chart of the frequencies would wrongly make that class look large.



The bars touch, since the data is continuous, and each bar starts and ends at the class boundaries.

Cumulative frequency graphs

The **cumulative frequency** at a value is how many data values are **less than or equal to** it: add up the frequencies as you go. Plot each running total at the **upper class boundary** of its class, start at the lower boundary of the first class with cumulative frequency 0, and join the points with a smooth curve.



To read the graph, go **across** from a cumulative frequency to the curve, then **down** to the value:

- median: across from $\frac{n}{2}$; quartiles: across from $\frac{n}{4}$ and $\frac{3n}{4}$ (here 50, 100 and 150, giving about 20, 25 and 31 minutes, so the IQR is about 11 minutes);

- the pth percentile: across from $p\%$ of n;
- how many are below a value: go **up** from the value and read the cumulative frequency; how many are **above** it is n minus that.

Choosing a diagram

Diagram	Advantages	Disadvantages
Stem-and-leaf	keeps every value; shows the shape; the mode, median, quartiles and mean can all be found from it	only for small data sets
Box-and-whisker	shows the median, quartiles and range at a glance; very easy to compare two sets; works for any size of data set	loses the individual values, so no mean or mode
Histogram	shows the shape of a large grouped data set; copes with unequal classes	the individual values are lost
Cumulative frequency graph	gives estimates of the median, quartiles and percentiles of large grouped data	the individual values are lost; estimates only

Key facts and formulas

"Given" means it is in the MF19 list of formulae in the exam. "Learn it" means you must remember it.

Formula	Status
Ungrouped: $\bar{x} = \dfrac{\sum x}{n}$, standard deviation $= \sqrt{\dfrac{\sum(x-\bar{x})^2}{n}} = \sqrt{\dfrac{\sum x^2}{n} - \bar{x}^2}$	Given
Grouped: $\bar{x} = \dfrac{\sum xf}{\sum f}$, standard deviation $= \sqrt{\dfrac{\sum(x-\bar{x})^2 f}{\sum f}} = \sqrt{\dfrac{\sum x^2 f}{\sum f} - \bar{x}^2}$, with x the midpoints	Given
Variance $=$ (standard deviation)^2	Learn it
Coded data: $\bar{x} = a + \dfrac{\sum(x-a)}{n}$; sd $= \sqrt{\dfrac{\sum(x-a)^2}{n} - \left(\dfrac{\sum(x-a)}{n}\right)^2}$	Learn it
$\sum(x-a) = \sum x - na$	Learn it
Back to totals: $\sum x = n\bar{x}$, $\sum x^2 = n(\sigma^2 + \bar{x}^2)$	Learn it
Median: the $\frac{n+1}{2}$th value; Q_1, Q_3: medians of the lower and upper halves	Learn it
Range $=$ largest $-$ smallest; IQR $= Q_3 - Q_1$	Learn it
Frequency density $= \dfrac{\text{frequency}}{\text{class width}}$	Learn it
Midpoint $= \frac{1}{2}$(lower boundary $+$ upper boundary)	Learn it
From a cumulative frequency graph: median at $\frac{n}{2}$, quartiles at $\frac{n}{4}$ and $\frac{3n}{4}$, pth percentile at $\frac{p}{100}n$	Learn it

How to do it

In the Paper 5 questions we have tagged for this topic (47 questions, 2021–2025), the types came up this often. 35 of the 47 questions mix two or more types, so the counts add up to more than 47.

Question type	Questions
Stem-and-leaf diagrams	16
Mean and standard deviation of grouped data	16
Comparing and interpreting data	15

Question type	Questions
Cumulative frequency graphs	13
Mean and standard deviation from totals	12
Histograms	10
Box-and-whisker plots	9
Which class holds the median or a quartile	6

1. Stem-and-leaf diagrams (16 questions)

Drawing a back-to-back diagram:

1. **Write the stems** in a column, in order, including any stem with no leaves. Don't split or reverse them.
2. **Put each set's leaves in order.** The left-hand set reads from **right to left**, smallest next to the stem; the right-hand set reads left to right. Space the leaves evenly so the columns line up, and use no commas.
3. **Label** both sides with the names from the question.
4. **Add one key** for both sets, with a real example, both names and the units, for example "$3 \mid 16 \mid 7$ means 163 cm for Team A and 167 cm for Team B".

Reading the median and IQR from one:

1. Count the values (n) and find the positions of Q_1, Q_2 and Q_3.
2. Count to each position. On the left-hand side, count outwards from the stem.
3. Turn each back into a real value using the key (for example $31 \mid 3$ in thousands of dollars is $31,300), then IQR $= Q_3 - Q_1$ with units.

Variations you will meet:

- **Decimals or large units in the key**, such as stems in seconds with leaves in tenths, or salaries in hundreds of dollars. Use the key every time you read a value.
- **"Find the missing value"** once the mean is known: total of all values $= n\times$ mean, so the missing value is that total minus the known values.
- **"State an advantage of a stem-and-leaf diagram over a box-and-whisker plot"**: it shows every data value, so you can still find the mean, the mode and the shape.

2. Mean and standard deviation of grouped data (16 questions)

1. **Find the frequencies.** If you are given cumulative frequencies, subtract neighbouring ones.
2. **Find each class midpoint** from the class boundaries (for "21–25 to the nearest minute", the midpoint is 23).
3. **Mean:** $\bar{x} = \dfrac{\sum xf}{\sum f}$. Write the whole sum out with your numbers in, then evaluate.
4. **Standard deviation:** $\sqrt{\dfrac{\sum x^2 f}{\sum f} - \bar{x}^2}$, again with the sum written out. Use the unrounded mean.

Variations you will meet:

- **The mean is given** and you only need the standard deviation: use the given mean for $\bar{x}$.
- **Exact answers:** if the mean comes out exactly, such as 57.375, give it exactly (or to 3 significant figures if it isn't exact).
- **"Without calculating, would the standard deviation increase, decrease or stay the same?"** after values are corrected: moving values towards the middle decreases it, moving them further out increases it.

3. Comparing and interpreting data (15 questions)

1. **Pick one measure of average and one of spread** for both sets, such as the medians and the IQRs.
2. **Write a sentence in context** for each comparison: "On average Team A were shorter than Team B (lower median)" and "Team B's heights were more consistent (smaller IQR)". Mentioning "average" or "spread" and the context earns the mark; just quoting two numbers does not.
3. For "make two comparisons", give **one about average and one about spread**.

Variations you will meet:

- **"Mean or median?"**: say which one and why, pointing to the data: "the median, because the value 95 is extreme and pulls the mean upwards", or "because the data is skewed".
- **Why the mean and median differ**: the distribution is skewed (or has an extreme value).
- **A quick raw-data question**: find the mean and median of a short list, then explain why one is more suitable.

4. Cumulative frequency graphs (13 questions)

Drawing:

1. **Make a running total** of the frequencies if the table gives plain frequencies.
2. **Plot each total at the upper class boundary** (9.5, 14.5, ... for classes recorded to the nearest whole number), plus a first point at the lower boundary of the first class with cumulative frequency 0.
3. **Join the points with a smooth increasing curve**, not straight line segments. Label the axes ("Cumulative frequency", "Time (t minutes)") and use linear scales with numbers marked, filling at least half the grid.

Reading:

1. Work out the cumulative frequency you need: $\frac{n}{2}$, $\frac{n}{4}$, $\frac{3n}{4}$ or $\frac{p}{100}n$.
2. Draw a line across to the curve and down to the axis, so the examiner can see it, and read the value.

Variations you will meet:

- **"30% took more than k minutes"**: then 70% took less, so go across from $0.7n$, not $0.3n$. Likewise, if 50 of 250 took more than k, go across from $250 - 50 = 200$.
- **"How many took between 15 and 30 minutes?"**: read the cumulative frequencies at both values and subtract.
- **"How many took more than T?"**: n minus the cumulative frequency at T.
- **Draw a box-and-whisker plot from the graph**: read the median and quartiles, and take the lowest and highest values from where the curve starts and ends.

5. Mean and standard deviation from totals (12 questions)

1. **Write down what you have**: n, and either $\sum x$ and $\sum x^2$, or $\sum(x-a)$ and $\sum(x-a)^2$.
2. **Coded totals:** mean $= a + \frac{\sum(x-a)}{n}$; standard deviation from the coded formula (no a in it).
3. **Unknown a or n:** use $\sum(x-a) = \sum x - na$. For example, if $n = 30$, $\sum(x-k) = 210$ and the mean is 25, then $\sum x = 750$, so $750 - 30k = 210$ and $k = 18$.
4. **Two sets together:** get n, $\sum x$ and $\sum x^2$ for each set, add them, then find the mean and standard deviation of the combined set. With coded totals using the same a, add the coded totals instead.

Variations you will meet:

- **Find $\sum x^2$ or $\sum y^2$ from a given standard deviation**: rearrange $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$.
- **Variance or standard deviation?** Read the question: variance has no square root.

6. Histograms (10 questions)

1. **Find the class boundaries and widths.** Watch out for "to the nearest minute" classes (1–10 has boundaries 0.5 and 10.5) and gaps between classes.
2. **Frequency density** $=$ **frequency** $\div$ **width** for every class. Write the densities down, for example in an extra row of the table.
3. **Draw the bars** from boundary to boundary with heights equal to the frequency densities, with a ruler and no gaps.
4. **Label the axes**: "Frequency density" and the variable with its units. Use linear scales with at least three values marked.

7. Box-and-whisker plots (9 questions)

1. **Find the five values**: lowest, Q_1, median, Q_3, highest (from a list, a stem-and-leaf diagram or a cumulative frequency graph).
2. **Draw one linear scale** with numbers marked, labelled with the variable and units, long enough for both plots.
3. **Draw each plot** above the scale: box from Q_1 to Q_3, line at the median, whiskers from the middle of the box ends to the lowest and highest values. Whiskers never go through the box.
4. **Label each plot** with its group's name.

8. Which class holds the median or a quartile (6 questions)

1. **Running totals** show where each value falls. With $n = 150$, the median is about the 75th value; if the running totals are $8, 20, 70, 118, \ldots$ then the 75th value is in the fourth class.
2. **Greatest possible IQR:** find the classes holding Q_1 and Q_3. The largest IQR is the **upper boundary** of the Q_3 class minus the **lower boundary** of the Q_1 class. ("Show that the IQR is not greater than 15" is the same idea.)

Worked examples

The bold codes show where marks are earned, as in Cambridge mark schemes. The number is how many marks.

- **M1** method mark: for a correct method, even if the numbers slip.
- **A1** accuracy mark: for a correct answer; it needs the M mark before it.
- **B1** mark for a correct result on its own, with no method needed.

Example 1

The resting heart rates, in beats per minute (bpm), of 15 members of a running club and 15 members of a chess club are:

Runners: 48, 52, 53, 55, 57, 58, 60, 61, 62, 64, 65, 66, 69, 71, 74

Chess club: 72, 58, 81, 66, 70, 90, 63, 74, 68, 61, 79, 72, 84, 67, 75

(a) Draw a back-to-back stem-and-leaf diagram, with the runners on the left. **[4]**

(b) Find the median and the interquartile range of the runners' heart rates. **[3]**

For the chess club, $Q_1 = 66$, the median is 72 and $Q_3 = 79$.

(c) Draw box-and-whisker plots for the two clubs in a single diagram. **[3]**

(d) Make one comparison between the two clubs. **[1]**

(a)

Runners	Stem	Chess club
8	4	
8 7 5 3 2	5	8
9 6 5 4 2 1 0	6	1 3 6 7 8
4 1	7	0 2 2 4 5 9
	8	1 4
	9	0

Key: $1 \mid 6 \mid 3$ means 61 bpm for a runner and 63 bpm for a chess club member.

B1 for the stems, **B1** for the runners' leaves in order outwards, **B1** for the chess club's leaves in order, **B1** for the key with both clubs and "bpm".

(b) $n = 15$: the median is the 8th value, 61 bpm. **B1**

Q_1 is the 4th value, 55, and Q_3 the 12th, 66. **M1**. IQR $= 66 - 55 = 11$ bpm. **A1**

(c) Runners: $48, 55, 61, 66, 74$. Chess club: $58, 66, 72, 79, 90$. Draw a scale from 40 to 90 labelled "Heart rate (bpm)", then both plots above it, each labelled. **B1** for each plot with all five values correct, **B1** for one labelled linear scale and whiskers drawn properly.

(d) On average the runners have lower resting heart rates than the chess club (median 61 bpm against 72 bpm). **B1** (Or: the runners' heart rates are less spread out, since their IQR is 11 against 13.)

Example 2

The masses of 120 eggs, to the nearest gram, are summarised below.

Mass (g)	40–49	50–54	55–59	60–64	65–74
Frequency	12	26	38	30	14

(a) Draw a histogram to represent this information. **[4]**

(b) Calculate an estimate of the mean mass. **[3]**

(c) State the class that contains the median. **[1]**

(a) The masses are to the nearest gram, so the boundaries are $39.5, 49.5, 54.5, 59.5, 64.5, 74.5$.

Class	39.5–49.5	49.5–54.5	54.5–59.5	59.5–64.5	64.5–74.5
Width	10	5	5	5	10
Frequency density	1.2	5.2	7.6	6.0	1.4

M1 for frequency ÷ width, **A1** for all heights drawn correctly, **B1** for the bars at the boundaries on a linear scale, **B1** for the axes labelled "Frequency density" and "Mass (g)".

(b) Midpoints: $44.5, 52, 57, 62, 69.5$. **B1**

$$\bar{x} = \frac{44.5 \times 12 + 52 \times 26 + 57 \times 38 + 62 \times 30 + 69.5 \times 14}{120} = \frac{6885}{120} = 57.375 \text{ g}.$$

M1 for the sum with midpoints, **A1** for 57.375 (or 57.4).

(c) The median is about the 60th value. The running totals are $12, 38, 76, \ldots$, so it is in the class 55–59 g. **B1**

Example 3

For 25 values of x, $\sum(x-50)=-40$ and $\sum(x-50)^2=1896$.

(a) Find the mean and standard deviation of these 25 values. **[3]**

(b) Find $\sum x$ and $\sum x^2$. **[2]**

A further 15 values of y have $\sum y=780$ and $\sum y^2=41\,100$.

(c) Find the mean and standard deviation of all 40 values. **[4]**

(a) Mean $=50+\dfrac{-40}{25}=50-1.6=48.4$. **B1**

$$\text{sd}=\sqrt{\frac{1896}{25}-(-1.6)^2}=\sqrt{75.84-2.56}=\sqrt{73.28}=8.56\ (3\text{ s.f.})$$

M1 A1

(b) $\sum x=\sum(x-50)+25\times 50=-40+1250=1210$. **B1**

$\sum x^2=25(\sigma^2+\bar{x}^2)=25(73.28+48.4^2)=25\times 2415.84=60\,396$. **B1**

(c) Mean $=\dfrac{1210+780}{40}=\dfrac{1990}{40}=49.75$. **M1 A1**

$$\text{sd}=\sqrt{\frac{60\,396+41\,100}{40}-49.75^2}=\sqrt{2537.4-2475.0625}=\sqrt{62.3375}=7.90\ (3\text{ s.f.})$$

M1 for the variance formula with the combined totals, **A1**.

Example 4

The times, t minutes, taken by 200 people to finish a task are summarised below.

Time (t minutes)	$0\leqslant t<10$	$10\leqslant t<20$	$20\leqslant t<25$	$25\leqslant t<30$	$30\leqslant t<40$	$40\leqslant t<70$
Frequency	14	36	50	44	38	18

(a) Draw a cumulative frequency graph. **[3]**

(b) Use your graph to estimate the median and the interquartile range. **[3]**

(c) 30% of the people took longer than k minutes. Estimate k. **[2]**

(d) Calculate estimates of the mean and standard deviation of the times. **[5]**

(a) Running totals at the upper boundaries: $(10,14),(20,50),(25,100),(30,144),(40,182),(70,200)$, starting from $(0,0)$. **B1** for the cumulative frequencies, **M1** for plotting at the upper boundaries, **A1** for a smooth curve with labelled axes. This is the fourth diagram in "Understand it".

(b) Median: across from 100, giving 25 minutes. **B1**

Q_1: across from 50, giving 20; Q_3: across from 150, giving about 31. **M1**. IQR $\approx 31-20=11$ minutes. **A1**

(c) 30% took **longer** than k, so 70% took less: $0.7\times 200=140$. **M1** Across from 140: $k\approx 29$ minutes. **A1**

(d) Midpoints: $5,15,22.5,27.5,35,55$. **B1**

$$\bar{x}=\frac{5\times 14+15\times 36+22.5\times 50+27.5\times 44+35\times 38+55\times 18}{200}=\frac{5265}{200}=26.325=26.3\text{ minutes }(3\text{ s.f.})$$

M1 A1

$$\text{variance} = \frac{5^2 \times 14 + 15^2 \times 36 + 22.5^2 \times 50 + 27.5^2 \times 44 + 35^2 \times 38 + 55^2 \times 18}{200} - 26.325^2 = \frac{168\,037.5}{200} - 693.0056 = 147.18$$

M1. Standard deviation $= \sqrt{147.18} = 12.1$ minutes. **A1**

Common mistakes

- **Drawing a bar chart instead of a histogram.** The heights must be frequency densities, not frequencies. Divide each frequency by the class width, and write the densities down before you draw.
- **Wrong class boundaries or widths.** "$11–20$ to the nearest minute" is 10.5 to 20.5, width 10, not 9. Mixing up boundaries shifts every bar and every midpoint.
- **Plotting a cumulative frequency at the wrong place.** Plot at the **upper boundary** of each class (for example 9.5, not 9 or the midpoint), start the curve at 0 at the lower boundary of the first class, not automatically at the origin, and draw a curve, not straight segments.
- **Reading a percentile the wrong way round.** "40% are k cm or more" means going across from 60% of n.
- **Reading the left side of a back-to-back diagram the wrong way.** Leaves on the left increase outwards from the stem, so $7\,5\,2\,|\,4$ means $42, 45, 47$.
- **A key that doesn't help.** Use one key with two different leaves, both group names and the units.
- **Using the wrong numbers for the mean.** Use midpoints, not upper boundaries, class widths or cumulative frequencies, and divide by the total frequency, not the number of classes.
- **Stopping at the variance**, or square-rooting when the variance was asked for. Read which one the question wants.
- **Rounding the mean before the standard deviation.** $\bar{x}^2$ is sensitive to rounding; keep the full value (or at least 4 significant figures) until the end.
- **Comparing without context.** "Set A has a larger IQR" earns nothing; "Team B's heights are more varied than Team A's" does.
- **Missing labels.** Every diagram needs labelled axes (with units), a scale with numbers, and every box plot needs a name.

How to get the marks

- **Diagrams are marked for accuracy.** Use a ruler, a sensible linear scale that fills at least half the grid, and at least three numbers on each scale. Label axes with the quantity and units; label each box plot.
- **Show the numbers behind a diagram**: frequency densities, cumulative frequencies, the five values for a box plot. A mark is often given for these even if the drawing slips.
- **Show the graph being used.** When you read from a cumulative frequency graph, draw the lines across and down so the examiner can see them.
- **Write out means and standard deviations in full** with your midpoints and frequencies in, not just the answer from a calculator. The method marks need to see the midpoints.
- **Identify every answer.** Write "median $=$", "IQR $=$", "sd $=$" so the right value is marked.
- **"Make a comparison"** means a sentence in context about average or spread. "Two comparisons" means one of each.
- **"Estimate"** is used for grouped data and graph readings. For graph readings a small range of answers is accepted, so read carefully to half a small square.
- **Accuracy:** non-exact answers to 3 significant figures. Keep 4 or more significant figures in working. If an answer is exact (such as 57.375) you may give it in full.

Quick check

1. The values $3, 7, 7, 8, 10, 12, 30$ are recorded. Find the mean, median and mode, and say whether the mean or the median is the better average here.
2. The distances, d km, travelled by 80 people are grouped as $0 \leqslant d < 5$: 10, $5 \leqslant d < 10$: 24, $10 \leqslant d < 20$: 30, $20 \leqslant d < 40$: 16. (a) Which class has the tallest bar in a histogram, and how tall is it? (b) Find the greatest possible interquartile range.

3. For n values of x, $\sum(x-10)=48$ and $\sum x=398$. Find n.
4. A group of 20 students has a mean test score of 64 and a standard deviation of 5. A 21st student, who scored 85, joins the group. Find the new mean and standard deviation.
5. Give one advantage of a stem-and-leaf diagram over a box-and-whisker plot, and one advantage of a box-and-whisker plot over a stem-and-leaf diagram.

Answers

1. Mean $= \frac{77}{7} = 11$; median $= 8$ (the 4th value); mode $= 7$. The median is better, because 30 is an extreme value that pulls the mean above almost all the data.
2. (a) Frequency densities: $10 \div 5 = 2$, $24 \div 5 = 4.8$, $30 \div 10 = 3$, $16 \div 20 = 0.8$. The tallest bar is $5 \leqslant d < 10$, height 4.8, even though $10 \leqslant d < 20$ has more people. (b) The running totals are $10, 34, 64, 80$. Q_1 (the 20th value) is in $5 \leqslant d < 10$ and Q_3 (the 60th) is in $10 \leqslant d < 20$, so the greatest possible IQR is $20 - 5 = 15$ km.
3. $\sum(x - 10) = \sum x - 10n$, so $398 - 10n = 48$ and $n = 35$.
4. Old total $= 20 \times 64 = 1280$, new mean $= \frac{1280+85}{21} = 65$. Old $\sum x^2 = 20(5^2 + 64^2) = 82\,420$, new $\sum x^2 = 82\,420 + 85^2 = 89\,645$. New sd $= \sqrt{\frac{89\,645}{21} - 65^2} = \sqrt{43.81} = 6.62$.
5. A stem-and-leaf diagram keeps every data value, so you can find the mean or mode from it. A box-and-whisker plot shows the median, quartiles and range clearly and works for large data sets, and two plots are easy to compare.

Past questions to try

- **9709/52/M/J/24 Q4**: median and IQR from a back-to-back stem-and-leaf diagram, box plots, and mean or median.
- **9709/51/M/J/24 Q3**: a histogram with unequal widths, then the greatest possible IQR.
- **9709/53/O/N/24 Q4**: a cumulative frequency graph, a percentile, then the grouped mean and standard deviation.
- **9709/53/M/J/21 Q3**: combining two teams' totals for the mean and standard deviation.