

Statistics

IGCSE Maths 0580 · Topic 9 revision notes

Everything from the 7 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. <u>Classifying and interpreting data</u> | 4. <u>Statistical charts and diagrams</u> |
| 2. <u>Averages, range and interquartile range</u> | 5. <u>Scatter diagrams and correlation</u> |
| 3. <u>Averages from frequency tables</u> | 6. <u>Cumulative frequency diagrams</u> |
| | 7. <u>Histograms with unequal class widths</u> |

1 Classifying and interpreting data

Syllabus 9.2

Classify first (qualitative or quantitative, discrete or continuous), count with a tally, and read every fraction from the right row or column. To compare two data sets, always compare one average AND one measure of spread, and say what each means in context.

Worked example: Classify the data

Classify each of these: (a) the colour of a car (b) the number of pages in a book (c) the time taken to run 100 m (d) a shoe size.

1. (a) Colours are words, not numbers: qualitative.
2. (b) Pages are counted, so only whole numbers are possible: quantitative, discrete.
3. (c) Time is measured and could be 12.3 s, 12.34 s, and so on: quantitative, continuous.
4. (d) Shoe sizes come only in fixed steps (whole and half sizes): quantitative, discrete.

Answer: (a) qualitative (b) discrete (c) continuous (d) discrete

Classify, count, read the right total, compare in context.

- ✓ **Classify:** words are qualitative; counts are discrete; measurements are continuous.
- ✓ **Tally:** one pass through the data, gates of five, then check the frequencies add to the total.
- ✓ **Two-way tables:** fill any row or column with only one gap, then check the last row or column.
- ✓ **Fractions:** the total after 'of' goes on the bottom.
- ✓ **Compare:** one average and one spread, each in context.

Watch out!

Always check that the frequencies add up to the total number of data values. A missing or double-counted value is the most common way to lose the mark for a frequency table.

Read the words after 'of'. 'What fraction of the girls...' divides by the girls' total, not by the grand total of 60. Using the wrong total is the examiner's favourite trap in two-way tables.

With times, a LOWER mean is better (faster). With spread, a smaller range means MORE consistent, not 'worse'. Always say what the numbers mean in context, not just 'the mean is bigger'.

2 Averages, range and interquartile range

Syllabus 9.3

Mean = $\frac{\text{sum of values}}{\text{number of values}}$; median = middle value once the data are IN ORDER, at position $\frac{n+1}{2}$; mode = most common; range = largest – smallest. (Extended)

Interquartile range = upper quartile – lower quartile.

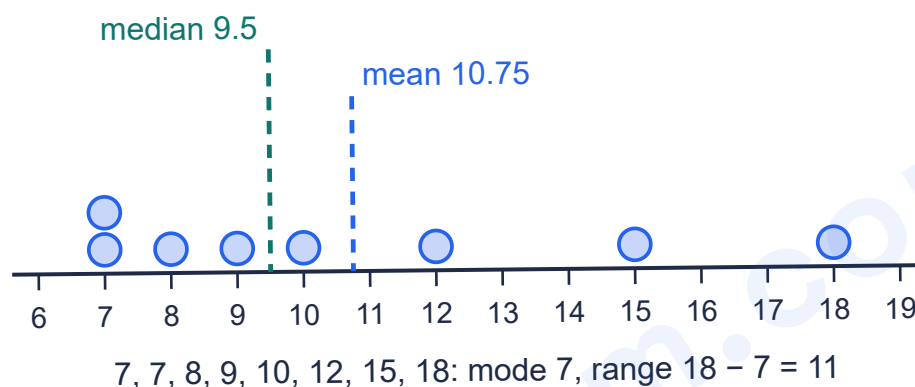


Fig. 1: the eight values 7, 7, 8, 9, 10, 12, 15, 18 on a dot plot, drawn to scale. The large value 18 pulls the mean (10.75) to the right of the median (9.5).

Worked example: All four measures

Find the mean, median, mode and range of 12, 7, 9, 15, 7, 10, 18, 8.

1. Order the data: 7, 7, 8, 9, 10, 12, 15, 18. There are $n = 8$ values.
2. Mean: $7 + 7 + 8 + 9 + 10 + 12 + 15 + 18 = 86$, and $86 \div 8 = 10.75$.
3. Median: position $\frac{8+1}{2} = 4.5$, halfway between the 4th and 5th values, 9 and 10, so 9.5.
4. Mode: 7 appears twice, every other value once.
5. Range: $18 - 7 = 11$.

Answer: Mean 10.75, median 9.5, mode 7, range 11 (Fig. 1)

Order first, then count positions.

- ✓ **Order** the data before finding the median or quartiles.
- ✓ **Mean:** $\text{total} \div \text{number of values}$; to find a missing value, work with totals ($\text{mean} \times n$).
- ✓ **Median** at position $\frac{n+1}{2}$; (Extended) quartiles at $\frac{n+1}{4}$ and $\frac{3(n+1)}{4}$.
- ✓ **Spread:** $\text{range} = \text{max} - \text{min}$; (Extended) $\text{IQR} = \text{UQ} - \text{LQ}$.
- ✓ **Choosing:** median when there is an outlier; mode for categories.

Watch out!

Order the data *BEFORE* finding the median. The middle of the unsorted list 12, 7, 9, 15, 7, 10, 18, 8 is 15 and 7, which gives a wrong 'median' of 11. And $\frac{n+1}{2}$ is the *POSITION* of the median, not its value.

'The mean, because it uses all the values' is not enough when there is an outlier. Look at the data: if one value is far from the rest, the median is the better choice, and you must say why.

The IQR is one number, $\text{UQ} - \text{LQ}$. Writing 'the IQR is 11 to 22' loses the mark, and so does finding quartiles from an unordered list.

3 Averages from frequency tables

Syllabus 9.3

Mean = $\frac{\sum fx}{\sum f}$: multiply each value by its frequency, add, then divide by the TOTAL frequency. For grouped data (Extended) use the midpoint of each class as x , and call the answer an estimate.

Worked example: All four from a table

The numbers of goals scored in 25 matches are shown.

Goals x 0 1 2 3 4

Frequency f 5 8 6 4 2

Find the mean, median, mode and range.

1. fx values: $0 \times 5 = 0$, $1 \times 8 = 8$, $2 \times 6 = 12$, $3 \times 4 = 12$, $4 \times 2 = 8$. So $\sum fx = 40$.

2. $\sum f = 25$, so mean = $\frac{40}{25} = 1.6$ goals.

3. Median position = $\frac{25 + 1}{2} = 13$. Running totals: 5 (up to 0 goals), 13 (up to 1 goal). The 13th value is 1.

4. Mode: the highest frequency is 8, for 1 goal. Range: $4 - 0 = 4$.

Answer: Mean 1.6, median 1, mode 1, range 4

Add an fx column, divide by the total frequency.

- ✓ **Mean:** $\frac{\sum fx}{\sum f}$.
- ✓ **Median:** position $\frac{n+1}{2}$, then count down the running totals.
- ✓ **Mode:** the value (or class) with the highest frequency.
- ✓ **Grouped (Extended):** use midpoints, and say 'estimate'.
- ✓ **Backwards:** $\text{total} = \text{mean} \times \text{number}$; add totals, never means.

Watch out!

Divide $\sum fx$ by the total FREQUENCY (25 matches), not by the number of rows (5). And the mode is the value 1, not its frequency 8.

Use the MIDPOINT, not the upper bound of each class. Using 10, 20, 30, 40 in the example gives 26 minutes, which is too big. Note also that the last class here is 20 wide, so its midpoint is 40, not 35.

The mean of the two means, $\frac{64 + 59}{2} = 61.5$, is wrong because the classes are different sizes. Combine TOTALS, then divide by the total number of students.

4 Statistical charts and diagrams

Syllabus 9.4

In a pie chart the whole circle, 360° , is the total frequency, so each item gets $\frac{360^\circ}{\text{total}}$. In a pictogram, always use the key. A stem-and-leaf diagram must be ordered and must have a key.

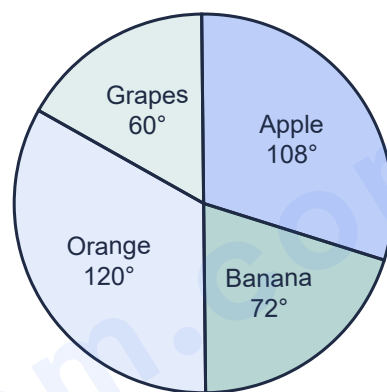
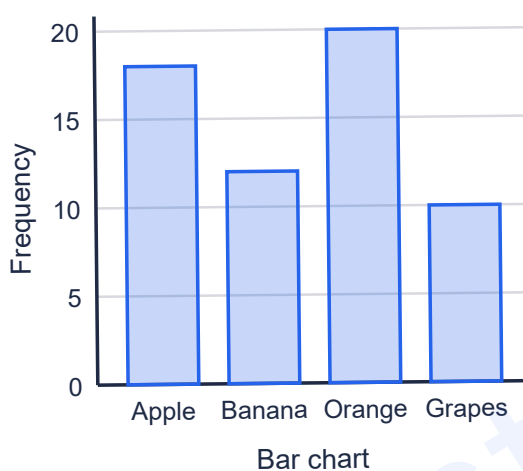


Fig. 1: the favourite fruit of 60 people (Apple 18, Banana 12, Orange 20, Grapes 10) as a bar chart and as a pie chart, both drawn to scale. Each person is worth $360^\circ \div 60 = 6^\circ$ of the pie.

Worked example: Read a pictogram

In a pictogram, one circle represents 4 books. Ali has 3 and a half circles, and Bea has 2 and a quarter circles. How many more books did Ali read than Bea?

1. Ali: $3.5 \times 4 = 14$ books.
2. Bea: $2.25 \times 4 = 9$ books.
3. Difference: $14 - 9 = 5$ books.

Answer: 5 books

Use the key, and use 360° for the whole.

- ✓ **Pie chart angles:** $\frac{360^\circ}{\text{total}}$ per item, then check the angles add to 360° .
- ✓ **Reading a pie chart:** $\text{number} = \frac{\text{angle}}{360} \times \text{total}$.
- ✓ **Pictograms:** $\text{symbols} \times \text{key value}$; part-symbols are fractions of it.
- ✓ **Composite bars:** $\text{section size} = \text{top reading} - \text{bottom reading}$.
- ✓ **Stem-and-leaf:** ordered leaves and a key, every time.

Watch out!

In a composite bar chart, read the size of each section, not the height of its top. The bus section that ends at 48 is $48 - 25 = 23$ students.

A bigger angle does not always mean more people when two pie charts have different totals. 80° of 72 people and 64° of 90 people are both 16 people.

No key, no marks: the examiner must know whether $5 \mid 1$ means 51, 5.1 or 510. And an unordered stem-and-leaf diagram also loses a mark.

5 Scatter diagrams and correlation

Syllabus 9.5

Positive correlation: as one variable increases, the other tends to increase. Negative: as one increases, the other tends to decrease. A line of best fit is ONE ruled line across the whole data set with roughly equal numbers of points on each side; use it to estimate only within the data.

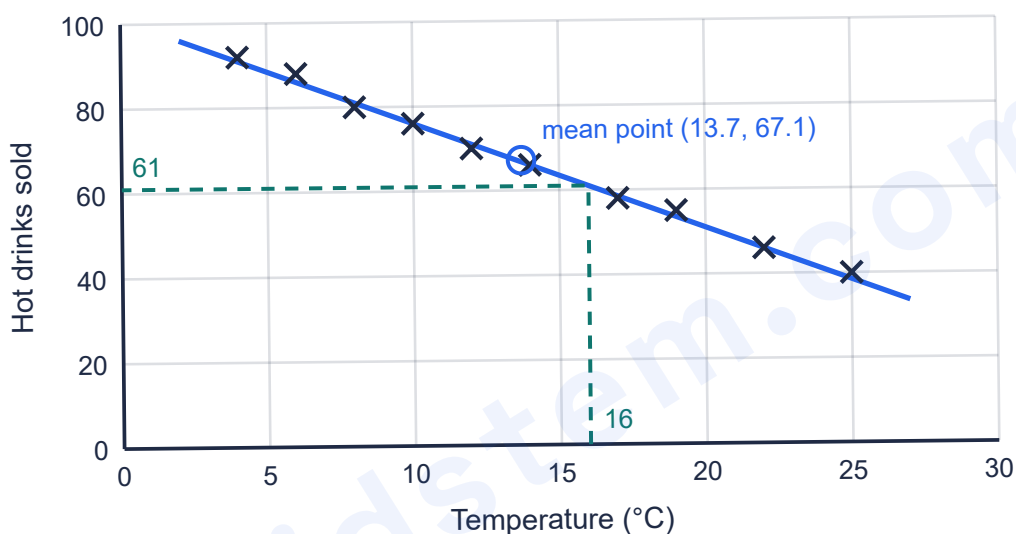


Fig. 1: temperature and hot drinks sold on 10 days, drawn to scale. The points show negative correlation. The line of best fit passes close to the mean point (13.7, 67.1) and gives an estimate of 61 drinks at 16°C.

Worked example: Describe and interpret

The table shows the temperature and the number of hot drinks sold at a cafe on 10 days.

Temperature (°C): 4, 6, 8, 10, 12, 14, 17, 19, 22, 25

Hot drinks: 92, 88, 80, 76, 70, 66, 58, 55, 46, 40

Describe the correlation and explain what it means.

1. Plot the 10 points (Fig. 1). They fall from left to right and lie close to a straight line.
2. So there is strong negative correlation.
3. In context: as the temperature increases, the number of hot drinks sold decreases.

Answer: Strong negative correlation: the warmer the day, the fewer hot drinks sold.

Describe, draw, then estimate inside the data.

- ✓ **Plot** small crosses; never join them.
- ✓ **Describe** the correlation (positive, negative, zero; strong or weak) and say what it means in context.
- ✓ **Draw** one ruled line across the data, near the mean point, with points evenly on each side.
- ✓ **Estimate** by reading up and across from the line.
- ✓ **Judge:** estimates outside the data are unreliable; correlation is not cause.

Watch out!

Correlation does not prove cause. Ice cream sales and sunburn cases have positive correlation, but ice cream does not cause sunburn: both rise in hot, sunny weather.

Do not join the points dot-to-dot, and do not force the line through the origin. Examiners also reject a line that stops short of the data or has most points on one side.

6 Cumulative frequency diagrams

Syllabus 9.6

Cumulative frequency is a running total. Plot each total at the UPPER boundary of its class and join with a smooth curve. For n values, read across at $\frac{n}{4}$, $\frac{n}{2}$ and $\frac{3n}{4}$ for the lower quartile, median and upper quartile.

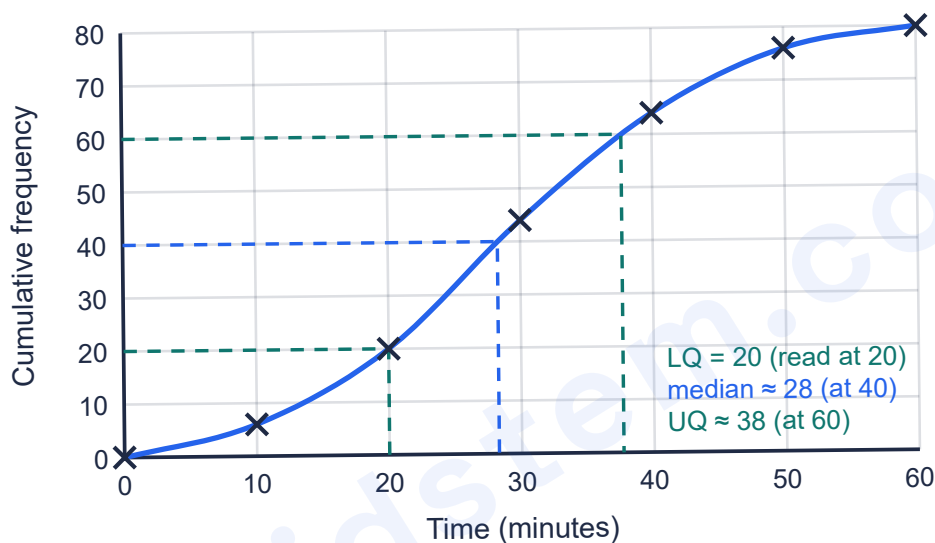


Fig. 1: cumulative frequency curve for the times of 80 people, drawn to scale from the table in the first worked example. Reading across at 20, 40 and 60 gives $LQ = 20$, median ≈ 28 and $UQ \approx 38$ minutes.

Worked example: From a frequency table to a curve

The times, t minutes, taken by 80 people to finish a puzzle are shown. Make a cumulative frequency table and list the points to plot.

Time	Frequency
$0 < t \leq 10$	6
$10 < t \leq 20$	14
$20 < t \leq 30$	24
$30 < t \leq 40$	20
$40 < t \leq 50$	12
$50 < t \leq 60$	4

1. Running totals: 6, $6 + 14 = 20$, $20 + 24 = 44$, $44 + 20 = 64$, $64 + 12 = 76$, $76 + 4 = 80$.
2. Check: the last total, 80, is the number of people.
3. Plot at the upper boundaries: (10, 6), (20, 20), (30, 44), (40, 64), (50, 76), (60, 80), starting from (0, 0).
4. Join the crosses with a smooth curve (Fig. 1).

Answer: cf: 6, 20, 44, 64, 76, 80; points (0, 0), (10, 6), (20, 20), (30, 44), (40, 64), (50, 76), (60, 80)

Running totals at upper boundaries.

- ✓ **Table:** add the frequencies as you go; the last total is n .
- ✓ **Plot** each total at the upper class boundary, as a cross, and join with a smooth curve.
- ✓ **Read** across at $\frac{n}{4}$, $\frac{n}{2}$, $\frac{3n}{4}$ (or $p\%$ of n), then down.
- ✓ **IQR** = $UQ - LQ$, both read from the horizontal axis.
- ✓ **More than:** total — reading.

Watch out!

Plotting at the midpoints (5, 15, 25, ...) shifts the whole curve 5 minutes to the left and every reading is wrong. The cf belongs to the upper boundary.

The IQR is the difference of two TIMES read from the horizontal axis ($38 - 20$), not the difference of the cumulative frequencies ($60 - 20 = 40$).

'How many took MORE than 45 minutes' is the total minus the reading. Writing 71, the number who took 45 minutes or less, is the most common lost mark on this topic.

7 Histograms with unequal class widths

Syllabus 9.7

Frequency density = $\frac{\text{frequency}}{\text{class width}}$. Plot frequency density on the vertical axis; then

frequency = frequency density $\times$ class width, which is the area of the bar.

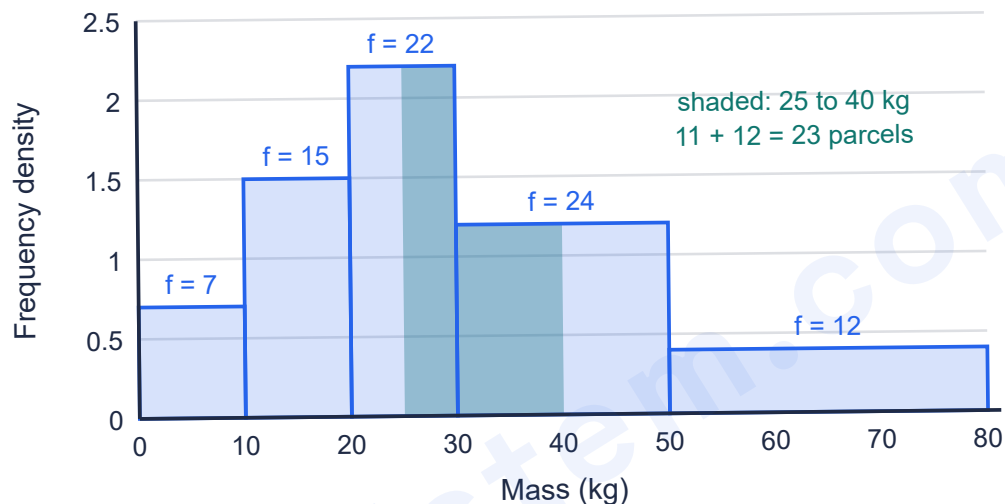


Fig. 1: histogram of the masses of 80 parcels, drawn to scale (frequency densities 0.7, 1.5, 2.2, 1.2, 0.4). The shaded area from 25 kg to 40 kg is $5 \times 2.2 + 10 \times 1.2 = 23$ parcels.

Worked example: Draw a histogram

The masses, m kg, of 80 parcels are shown. Calculate the frequency densities for a histogram.

Mass	Frequency
$0 < m \leq 10$	7
$10 < m \leq 20$	15
$20 < m \leq 30$	22
$30 < m \leq 50$	24
$50 < m \leq 80$	12

1. Class widths: 10, 10, 10, 20, 30.
2. Frequency densities: $7 \div 10 = 0.7$, $15 \div 10 = 1.5$, $22 \div 10 = 2.2$, $24 \div 20 = 1.2$, $12 \div 30 = 0.4$.
3. Draw bars from 0 to 10, 10 to 20, ..., 50 to 80 with these heights (Fig. 1).

Answer: 0.7, 1.5, 2.2, 1.2, 0.4

Divide to draw, multiply to read.

- ✓ **Width** = upper — lower boundary of each class.
- ✓ **Draw:** frequency density = frequency $\div$ width.
- ✓ **Read:** frequency = frequency density $\times$ width (the area).
- ✓ **Part of a class:** use only the width inside your range.
- ✓ **Hidden scale:** use a known class to find what one unit of height is worth.

Watch out!

The 30 to 50 class has the biggest frequency (24) but NOT the tallest bar: its density is only 1.2. Plotting frequency instead of frequency density is the mistake examiners look for first.

Using the whole class when only part is wanted overcounts: 25 to 40 kg is NOT $22 + 24 = 46$. Use only the widths 5 and 10 inside the range.

Comparing bar heights alone gives $\frac{6}{4.5} \times 30 = 40$, which ignores that the second class is only half as wide. Always multiply by the class width.