

YEAR 1 STATISTICS • CHAPTER 2

Measures of Location and Spread

Specification: Edexcel Paper 3 §2.3 — Measures of central tendency and variation • AQA section L3

KEY FACTS — THIS CHAPTER

- **Mean** $\bar{x} = \frac{\sum x}{n}$, or $\frac{\sum fx}{\sum f}$ from a frequency table. The **median** is the middle value; the **mode** is the most common.
- **Quartiles** split ordered data into four equal parts. **IQR** = $Q_3 - Q_1$ is the spread of the middle half.
- For **grouped** data, estimate the mean with class midpoints, and estimate the median, quartiles and percentiles by **linear interpolation**.
- **Variance** $\sigma^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2$. The **standard deviation** σ is its square root: the typical distance from the mean.
- **Coding** $y = \frac{x - a}{b}$: then $\bar{y} = \frac{\bar{x} - a}{b}$ and $\sigma_y = \frac{\sigma_x}{b}$. Adding or subtracting moves the mean but not the spread.
- When comparing data sets, compare a measure of **location** and a measure of **spread, in context**.

Two numbers summarise most data sets: one for where the values sit (location) and one for how widely they spread. This chapter builds the standard choices for each, shows how to get them from raw data, frequency tables and grouped data, and ends with the skill examiners reward most: saying what they mean.

1. Averages

KEY RESULT

$$\bar{x} = \frac{\sum x}{n} \quad \text{or, from a frequency table,} \quad \bar{x} = \frac{\sum fx}{\sum f}$$

The **median** is the middle value of the ordered data. The **mode** is the most frequent value (or the **modal class** for grouped data).

Each average has its strengths:

Average	Use it when...	Watch out for...
mean	you want every value to count; no extreme values	it is pulled towards outliers
median	there are outliers or the data are skewed	it ignores the size of extreme values
mode	the data are qualitative, or you want the most common value	there may be no mode, or several

For **grouped** data you don't know the exact values, so each class is represented by its **midpoint** and the mean is an **estimate**.

2. Quartiles, percentiles and the IQR

The **lower quartile** Q_1 , **median** Q_2 and **upper quartile** Q_3 cut the ordered data into four equal parts. The **interquartile range** $IQR = Q_3 - Q_1$ measures the spread of the middle 50%, so it ignores the extremes at both ends. **Percentiles** divide the data into 100 parts, and the 10th to 90th percentile range is another spread measure that is robust to extremes.

QUARTILES OF RAW (DISCRETE) DATA

For n ordered values, to find Q_1 work out $\frac{n}{4}$:

- if it is **not** a whole number, round **up** and take that value;
- if it **is** a whole number, take the mean of that value and the next one.

Use $\frac{n}{2}$ for the median and $\frac{3n}{4}$ for Q_3 in the same way.

WORKED EXAMPLE 2.1 · AVERAGES AND QUANTILES FROM RAW DATA

The times, in minutes, that 11 people took to complete a puzzle were 12, 15, 9, 22, 17, 15, 11, 19, 14, 15, 20. Find the mean, median, mode, quartiles and interquartile range.

In order: 9, 11, 12, 14, 15, 15, 15, 17, 19, 20, 22.

- Mean: $\frac{169}{11} = 15.4$ minutes (3 s.f.). Mode: 15 minutes.
- $\frac{n}{2} = 5.5$, so round up: the median is the 6th value, 15 minutes.
- $\frac{n}{4} = 2.75$, so Q_1 is the 3rd value, 12. $\frac{3n}{4} = 8.25$, so Q_3 is the 9th value, 19.
- IQR = $19 - 12 = 7$ minutes.

Grouped data: linear interpolation

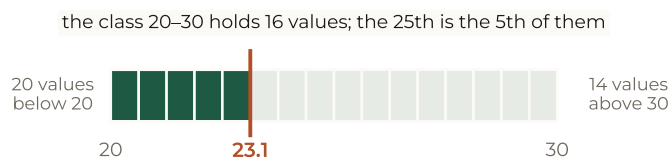
In grouped continuous data the values inside a class are unknown, so assume they are **spread evenly** through it. Then a value's position within its class gives its size.

LINEAR INTERPOLATION

To estimate the value at position k (use $k = \frac{n}{2}$ for the median, $\frac{n}{4}$ for Q_1 , $\frac{3n}{4}$ for Q_3 , $\frac{pn}{100}$ for the p th percentile):

$$\text{value} \approx L + \frac{k - C}{f} \times w$$

where L is the lower boundary of the class containing position k , C is the cumulative frequency before that class, f is its frequency and w is its width. (Named in Edexcel's guidance, and the standard method for both boards.)



Interpolation assumes the 16 values in 20–30 are evenly spread. The 25th value overall is the 5th of them: $20 + 5/16 \times 10 \approx 23.1$.

WORKED EXAMPLE 2.2 · ESTIMATES FROM GROUPED DATA

The table shows the journey times of 50 students. Estimate the mean, the median, the quartiles and the 90th percentile.

Time, t (minutes)	$0 \leq t < 10$	$10 \leq t < 20$	$20 \leq t < 30$	$30 \leq t < 40$	$40 \leq t < 60$
Frequency, f	6	14	16	10	4
Midpoint, x	5	15	25	35	50
Cumulative frequency	6	20	36	46	50

Journey times of 50 students

Mean: $\sum fx = 30 + 210 + 400 + 350 + 200 = 1190$, so $\bar{x} \approx \frac{1190}{50} = 23.8$ minutes.

Median: $k = 25$, which lies in $20 \leq t < 30$ (20 values come before it):

$$Q_2 \approx 20 + \frac{25 - 20}{16} \times 10 = 23.1 \text{ minutes (3 s.f.)}$$

Quartiles: Q_1 ($k = 12.5$) $\approx 10 + \frac{6.5}{14} \times 10 = 14.6$ and Q_3 ($k = 37.5$) $\approx 30 + \frac{1.5}{10} \times 10 = 31.5$, so the IQR is about 16.9 minutes.

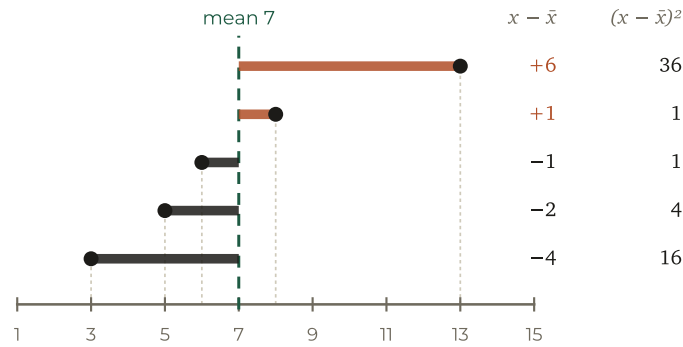
90th percentile: $k = 45$, in $30 \leq t < 40$: $\approx 30 + \frac{45 - 36}{10} \times 10 = 39$ minutes.

COMMON MISTAKE

For grouped continuous data use $\frac{n}{2}$, not $\frac{n+1}{2}$, and don't round positions: interpolation handles fractions. The "round up" rule is only for raw discrete data.

3. Variance and standard deviation

The IQR ignores most of the data. The **standard deviation** uses every value: it measures how far, typically, the values sit from the mean.



Deviations from the mean always sum to 0, so square them first. Here the mean square is $58 \div 5 = 11.6$, and $\sigma = \sqrt{11.6} \approx 3.41$.

VARIANCE AND STANDARD DEVIATION

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2, \quad \sigma = \sqrt{\sigma^2}$$

From a frequency table: $\sigma^2 = \frac{\sum fx^2}{\sum f} - \bar{x}^2$.

The second form of the variance, “the mean of the squares minus the square of the mean”, is the one to calculate with. It only needs n , $\sum x$ and $\sum x^2$, which questions often give you instead of the raw data (the **summary statistics**).

WORKED EXAMPLE 2.3 • FROM SUMMARY STATISTICS

For 20 values, $\sum x = 364$ and $\sum x^2 = 6964$. Find the mean and the standard deviation.

$$\bar{x} = \frac{364}{20} = 18.2$$

$$\sigma^2 = \frac{6964}{20} - 18.2^2 = 348.2 - 331.24 = 16.96, \quad \sigma = 4.12 \text{ (3 s.f.)}$$

WORKED EXAMPLE 2.4 · FROM A FREQUENCY TABLE

The numbers of goals scored in 30 matches are shown below. Find the mean, median and standard deviation.

Goals, x	0	1	2	3	4	5
Frequency, f	4	8	9	5	3	1

$$\sum f = 30, \sum fx = 58 \text{ and } \sum fx^2 = 0 + 8 + 36 + 45 + 48 + 25 = 162.$$

$$\bar{x} = \frac{58}{30} = 1.93 \text{ goals (3 s.f.)}$$

The median is the mean of the 15th and 16th values. The cumulative frequencies are 4, 12, 21, ..., so both are 2: the median is 2 goals.

$$\sigma^2 = \frac{162}{30} - \left(\frac{58}{30}\right)^2 = 1.662\dots, \quad \sigma = 1.29 \text{ goals (3 s.f.)}$$

ON YOUR CALCULATOR

The **1-variable statistics** mode gives $\bar{x}$, $\sum x$, $\sum x^2$ and two standard deviations. σ_x divides by n , as in this chapter. s_x divides by $n - 1$: AQA accepts either, but Edexcel expects σ_x . Enter frequency tables with the frequency column switched on.

4. Coding

Awkward numbers can be **coded** to simpler ones, for example $y = \frac{x - 150}{5}$. Work with y , then decode. (Named in Edexcel's guidance.)

CODING

If $y = \frac{x - a}{b}$, then

$$\bar{y} = \frac{\bar{x} - a}{b} \quad \text{and} \quad \sigma_y = \frac{\sigma_x}{b}$$

Subtracting a shifts every value, and so the mean, but leaves the spread alone. Dividing by b scales both.

WORKED EXAMPLE 2.5 · DECODING

Heights x cm were coded using $y = \frac{x - 150}{5}$, giving $\bar{y} = 2.4$ and $\sigma_y = 1.6$. Find the mean and standard deviation of the heights.

Rearrange: $x = 5y + 150$. So $\bar{x} = 5(2.4) + 150 = 162$ cm and $\sigma_x = 5(1.6) = 8$ cm. The $+150$ affects the mean only.

5. Comparing distributions**KEY RESULT**

To compare two data sets, make **two** comparisons, each **in context**:

1. a measure of **location**: mean or median;
2. a measure of **spread**: standard deviation, IQR or range.

Pair them sensibly: mean with standard deviation, or median with IQR (the median and IQR are better if there are outliers).

IN CONTEXT

“The mean time for Class A (15.4 min) was lower than for Class B (18.1 min), so on average Class A completed the puzzle faster. Class A’s standard deviation (3.75 min) was also smaller than Class B’s (6.2 min), so their times were more consistent.”

COMMON MISTAKE

“Class A has a lower mean” earns nothing on its own. Say what it means for the people or things in the question: “on average, Class A were faster”. The context sentence carries the marks.

Practice

Twelve questions in three tiers. Try each one properly before opening the solution. Give answers to 3 significant figures where they are not exact, and use the divisor n for variance and standard deviation.

Warm-up**Question 1**

[3 marks]

Ten students scored: 13, 7, 21, 4, 10, 15, 7, 18, 8, 12.

Find the mean, the median and the mode.

SOLUTION

$$\text{Mean} = \frac{115}{10} = 11.5.$$

In order: 4, 7, 7, 8, 10, 12, 13, 15, 18, 21. Here $\frac{n}{2} = 5$, a whole number, so the median is the mean of the 5th and 6th values: $\frac{10 + 12}{2} = 11$.

Mode = 7.

Question 2

[4 marks]

Twelve values, in order, are 3, 5, 6, 8, 9, 11, 12, 14, 15, 17, 20, 24. Find the median, the quartiles and the interquartile range.

SOLUTION

$$\frac{n}{4} = 3, \text{ so } Q_1 = \frac{6 + 8}{2} = 7.$$

$$\frac{n}{2} = 6, \text{ so } Q_2 = \frac{11 + 12}{2} = 11.5.$$

$$\frac{3n}{4} = 9, \text{ so } Q_3 = \frac{15 + 17}{2} = 16.$$

$$\text{IQR} = 16 - 7 = 9.$$

Question 3

[4 marks]

The table shows the number of pets owned by 25 students.

Pets, x	0	1	2	3
Frequency, f	7	10	6	2

Find the mean, the median and the mode.

SOLUTION

$\sum fx = 0 + 10 + 12 + 6 = 28$, so the mean is $\frac{28}{25} = 1.12$ pets.

$\frac{n}{2} = 12.5$, so round up: the median is the 13th value. The cumulative frequencies are 7, 17, ..., so the 13th value is 1 pet.

The mode is 1 pet (the highest frequency, 10).

Question 4

[3 marks]

Find the mean and standard deviation of 2, 4, 4, 4, 5, 5, 7, 9.

SOLUTION

$\sum x = 40$, so $\bar{x} = \frac{40}{8} = 5$.

$\sum x^2 = 4 + 16 + 16 + 16 + 25 + 25 + 49 + 81 = 232$, so

$$\sigma^2 = \frac{232}{8} - 5^2 = 29 - 25 = 4, \quad \sigma = 2$$

Standard**Question 5**

[6 marks]

For a set of 15 values, $\sum x = 270$ and $\sum x^2 = 5160$.

- Find the mean and the standard deviation.
- A 16th value, 18, is added to the set. Find the new mean and standard deviation.
- Explain why the standard deviation changed as it did.

SOLUTION

(a) $\bar{x} = \frac{270}{15} = 18$ and $\sigma^2 = \frac{5160}{15} - 18^2 = 344 - 324 = 20$, so $\sigma = 4.47$ (3 s.f.).

(b) Now $n = 16$, $\sum x = 288$ and $\sum x^2 = 5160 + 324 = 5484$.

$\bar{x} = \frac{288}{16} = 18$ and $\sigma^2 = \frac{5484}{16} - 18^2 = 342.75 - 324 = 18.75$, so $\sigma = 4.33$ (3 s.f.).

(c) The new value equals the mean, so the mean is unchanged, but it adds a deviation of zero to the set. The average squared deviation therefore falls, and the standard deviation with it.

Question 6

[6 marks]

The heights, h cm, of 40 seedlings are shown below.

Height, h (cm)	$0 \leq h < 5$	$5 \leq h < 10$	$10 \leq h < 15$	$15 \leq h < 20$	$20 \leq h < 30$
Frequency	4	10	15	8	3

Heights of 40 seedlings

- (a) Write down the modal class.
 (b) Estimate the mean height.
 (c) Estimate the median height.

SOLUTION

(a) $10 \leq h < 15$ (frequency 15).

(b) Midpoints 2.5, 7.5, 12.5, 17.5, 25:

$$\sum fx = 10 + 75 + 187.5 + 140 + 75 = 487.5$$

$$\bar{h} \approx \frac{487.5}{40} = 12.2 \text{ cm (3 s.f.)}$$

(c) $k = 20$. The cumulative frequencies are 4, 14, 29, so the median is in $10 \leq h < 15$:

$$Q_2 \approx 10 + \frac{20 - 14}{15} \times 5 = 12 \text{ cm}$$

Question 7

[5 marks]

The values x were coded using $y = \frac{x - 20}{4}$. For the 10 coded values, $\sum y = 30$ and $\sum y^2 = 190$.

- (a) Find the mean and standard deviation of y .
 (b) Hence find the mean and standard deviation of x .

SOLUTION

(a) $\bar{y} = \frac{30}{10} = 3$ and $\sigma_y^2 = \frac{190}{10} - 9 = 10$, so $\sigma_y = 3.16$ (3 s.f.).

(b) $x = 4y + 20$, so $\bar{x} = 4(3) + 20 = 32$ and $\sigma_x = 4\sqrt{10} = 12.6$ (3 s.f.).

Question 8

[5 marks]

Using the seedling data from Question 6:

- (a) Estimate the lower and upper quartiles, and the interquartile range.
- (b) Estimate the 90th percentile.

SOLUTION

Cumulative frequencies: 4, 14, 29, 37, 40.

(a) Q_1 : $k = 10$, in $5 \leq h < 10$, so $Q_1 \approx 5 + \frac{10 - 4}{10} \times 5 = 8$ cm.

Q_3 : $k = 30$, in $15 \leq h < 20$, so $Q_3 \approx 15 + \frac{30 - 29}{8} \times 5 = 15.6$ cm (3 s.f.).

$IQR \approx 15.625 - 8 = 7.63$ cm (3 s.f.).

(b) $k = 0.9 \times 40 = 36$, in $15 \leq h < 20$: $\approx 15 + \frac{36 - 29}{8} \times 5 = 19.4$ cm (3 s.f.).

Exam style**Question 9**

[6 marks]

Two classes sat the same test. Class A had mean 62 marks, standard deviation 8.5, median 64 and IQR 10. Class B had mean 58 marks, standard deviation 14, median 62 and IQR 21.

- (a) Compare the two classes' results.
- (b) Class B's teacher says the class contains a few very low marks. Explain how the given statistics support this, and say which pair of measures is the better summary for Class B.

SOLUTION

(a) Class A's mean (62) is higher than Class B's (58), so on average Class A did better on the test. Class A's standard deviation (8.5) is much smaller than Class B's (14), so Class A's marks were far more consistent, while Class B's were widely spread.

(b) For Class B the mean (58) is **below** the median (62). A few very low marks drag the mean down while barely moving the median, which is exactly the pattern here, and they also inflate the standard deviation (14) and the IQR (21). Since extreme values distort the mean and standard deviation, the **median and IQR** are the better summary for Class B.

Question 10

[8 marks]

The waiting times, t minutes, of 80 patients at a clinic are recorded.

Time, t (min)	$0 \leq t < 5$	$5 \leq t < 10$	$10 \leq t < 15$	$15 \leq t < 25$	$25 \leq t < 40$
Frequency	12	25	20	15	8

Waiting times of 80 patients

- Estimate the mean and the standard deviation.
- Estimate the median.
- Explain why your answers are estimates.

SOLUTION

(a) Midpoints 2.5, 7.5, 12.5, 20, 32.5.

$$\sum ft = 30 + 187.5 + 250 + 300 + 260 = 1027.5, \text{ so } \bar{t} \approx \frac{1027.5}{80} = 12.8 \text{ min (3 s.f.)}$$

$$\sum ft^2 = 75 + 1406.25 + 3125 + 6000 + 8450 = 19056.25, \text{ so}$$

$$\sigma^2 \approx \frac{19056.25}{80} - 12.84375^2 = 73.24 \dots, \quad \sigma \approx 8.56 \text{ min (3 s.f.)}$$

(b) $k = 40$. Cumulative frequencies are 12, 37, 57, so the median lies in $10 \leq t < 15$:

$$Q_2 \approx 10 + \frac{40 - 37}{20} \times 5 = 10.75 \text{ min}$$

(c) The original values are lost once the data are grouped. The calculations assume every value sits at its class midpoint (for the mean) or that values are evenly spread through the class (for the median), and neither is exactly true.

Question 11

[6 marks]

Group 1 has 12 values with mean 15 and standard deviation 2. Group 2 has 8 values with mean 20 and standard deviation 3. Find the mean and standard deviation of all 20 values combined.

SOLUTION

Mean: $\sum x = 12(15) + 8(20) = 180 + 160 = 340$, so $\bar{x} = \frac{340}{20} = 17$.

Sums of squares: from $\sigma^2 = \frac{\sum x^2}{n} - \bar{x}^2$ we get $\sum x^2 = n(\sigma^2 + \bar{x}^2)$:

$$12(4 + 225) = 2748, \quad 8(9 + 400) = 3272, \quad \text{total } 6020$$

$$\sigma^2 = \frac{6020}{20} - 17^2 = 301 - 289 = 12, \quad \sigma = 3.46 \text{ (3 s.f.)}$$

IN CONTEXT

The combined standard deviation (3.46) is bigger than either group's, because the two groups are centred on different means: that gap adds to the overall spread.

Question 12

[6 marks]

For 25 values a student calculates a mean of 40 and a standard deviation of 6. She then notices that one value was recorded as 52 when it should have been 42.

- (a) Find the correct mean.
- (b) Find the correct standard deviation.

SOLUTION

(a) From the wrong data, $\sum x = 25 \times 40 = 1000$. Correcting the value: $\sum x = 1000 - 52 + 42 = 990$, so the mean is $\frac{990}{25} = 39.6$.

(b) From $\sum x^2 = n(\sigma^2 + \bar{x}^2) = 25(36 + 1600) = 40900$. Correcting:

$$\sum x^2 = 40900 - 52^2 + 42^2 = 40900 - 2704 + 1764 = 39960$$

$$\sigma^2 = \frac{39960}{25} - 39.6^2 = 1598.4 - 1568.16 = 30.24$$

$$\sigma = 5.50 \text{ (3 s.f.)}$$

IN CONTEXT

Correcting one value that was too high pulls both the mean and the spread down.