Central Tendency and Variation

The mean, median and mode are measures of central tendency or location (roughly speaking, where the centre of the data lies). Then there's the variance and standard deviation, which measure variation and- hey, don't fall asleep...

The **Definitions** are really GCSE stuff

You probably already know these measures of **central tendency**, so learn them now — you'll be needing them loads.

Mean = $\bar{x} = \frac{\Sigma x}{n}$ or $\frac{\Sigma fx}{\Sigma f}$ where each x is a **data value**, f is the **frequency** of each x (the number of times it occurs), and n is the **total number** of data values.

Σ (sigma) means 'add stuff up' — so Σx means 'add up all the values of x'.

Median = **middle** data value when all the data values are placed **in order of size**.

Mode = **most frequently occurring** data value.

If n is **even**, the **median** is the **average of the middle two values** (the $\frac{n}{2}$th and the $\left(\frac{n}{2}+1\right)$th values).

If n is **odd**, the **median** is the **middle value** (round up $\frac{n}{2}$ to find its position).

Use a **Table** when there are a lot of **Numbers**

Example: The number of letters received one day in 100 houses was recorded. Find the mean, median and mode of the number of letters.

No. of letters	No. of houses
0	11
1	25
2	27
3	21
4	9
5	7

The first thing to do is make a **table** like this one:

No. of letters (x)	No. of houses (f)	fx
0	11 (11)	0
1	25 (36)	25
2	27 (63)	54
3	21	63
4	9	36
5	7	35
Totals	100	213

Multiply x by f to get this column.

Put the running total in brackets — it's handy when you're finding the median (but you can stop when you get past halfway).

$\Sigma f = 100$

$\Sigma fx = 213$

The number of letters received by each house is a discrete quantity (e.g. 3 letters). There isn't a continuous set of possible values between getting 3 and 4 letters (e.g. 3.45 letters).

1. Use the totals of the columns to find the mean: **mean** = $\frac{\Sigma fx}{\Sigma f} = \frac{213}{100}$ = **2.13 letters**
2. $\frac{n}{2} = \frac{100}{2} = 50$, so the median is **halfway between** the 50th and 51st data values.
   The **running total** of f shows that the data values in positions 37 to 63 are all 2s.
   This includes positions 50 and 51, so the **median = 2 letters**
3. The **highest frequency** is for 2 letters — so the **mode = 2 letters**

The mean number of letters received increased after this couple moved into the street.

The **Standard Deviation Formulas** look pretty **Tricky**

Standard deviation and **variance** both measure **variation** — i.e. how **spread out** the data is from the mean. The bigger the variance, the more spread out your readings are.

Variance = $\frac{\sum(x-\bar{x})^2}{n}$ or $\frac{\sum x^2}{n} - \bar{x}^2$ or $\frac{\sum fx^2}{\sum f} - \bar{x}^2$

Standard deviation = $\sqrt{\text{variance}}$

The x-values are the data, $\bar{x}$ is the mean, f is the frequency of each x, and n is the number of data values.

Make sure you're comfortable with these formulas and all the versions on the formula sheet (see p.257).

You might see S_{xx} used instead of $\Sigma(x - \bar{x})^2$.

Example: Find the mean and standard deviation of the following numbers: 2, 3, 4, 4, 6, 11, 12

Find the total of the numbers first: $\Sigma x = 2 + 3 + 4 + 4 + 6 + 11 + 12 = 42$

Then the mean is easy: mean = $\bar{x} = \frac{\Sigma x}{n} = \frac{42}{7} = 6$

Next find the sum of the squares: $\Sigma x^2 = 4 + 9 + 16 + 16 + 36 + 121 + 144 = 346$

Use this to find the variance: Variance = $\frac{\Sigma x^2}{n} - \bar{x}^2 = \frac{346}{7} - 6^2 = 49.428... - 36 = 13.428...$

Take the square root to find the standard deviation: Standard deviation = $\sqrt{13.428...} = 3.66$ (3 s.f.)

Central Tendency and Variation

Questions about **Standard Deviation** can look a bit **Weird**

They can ask questions about standard deviation in different ways. But you just need to use the same old formulas.

Example: The mean of 10 boys' heights is 180 cm, and the standard deviation is 10 cm. The mean for 9 girls is 165 cm, and the standard deviation is 8 cm. Find the mean and standard deviation of the whole group of 19 girls and boys.

Let the boys' heights be x and the girls' heights be y.

Write down the formula for the mean and put the numbers in for the boys: $\bar{x} = \frac{\Sigma x}{n} \Rightarrow 180 = \frac{\Sigma x}{10} \Rightarrow \Sigma x = 1800$

Do the same for the girls: $165 = \frac{\Sigma y}{9} \Rightarrow \Sigma y = 1485$

So the sum of the heights for the boys and the girls = $\Sigma x + \Sigma y = 1800 + 1485 = 3285$

And the **mean height** of the boys and the girls is: $\frac{3285}{19} =$ **172.9 cm** (1 d.p.)

Round the fraction to give your answer. But if you need to use the mean in more calculations, use the fraction (or your calculator's memory) so you don't lose accuracy.

Now for the **variance**.

Write down the formula for the boys first:

$$10^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 \Rightarrow 10^2 = \frac{\Sigma x^2}{10} - 180^2 \Rightarrow \Sigma x^2 = 10 \times (100 + 32\,400) = 325\,000$$

Do the same for the girls:

$$8^2 = \frac{\Sigma y^2}{n} - \bar{y}^2 \Rightarrow 8^2 = \frac{\Sigma y^2}{9} - 165^2 \Rightarrow \Sigma y^2 = 9 \times (64 + 27\,225) = 245\,601$$

So the sum of the squares of the heights of the boys and the girls is: $\Sigma x^2 + \Sigma y^2 = 325\,000 + 245\,601 = 570\,601$

Don't use the rounded mean (172.9) — you'll lose accuracy.

The **variance** of all the heights is: $\frac{570\,601}{19} - \left(\frac{3285}{19}\right)^2 = 139.041\ldots \text{ cm}^2$

The **standard deviation** of all the heights is: $\sqrt{139.041\ldots} =$ **11.8 cm** (3 s.f.)

The standard deviation uses the same units as the mean, so the units for the variance must be squared.

Practice Questions

Q1 Calculate the mean, median and mode of the data in the table on the right.

x	0	1	2	3	4
f	5	4	4	2	1

Q2 Find the mean and standard deviation of the following numbers:
11, 12, 14, 17, 21, 23, 27

Q3 The scores from 50 reviews of a product are recorded in the table below.

Score	1	2	3	4	5
Frequency	6	11	22	9	2

Calculate the mean and variance of the data.

For Q3, add rows for fx and fx^2. Then use the third variance formula given on the previous page.

Q4 a) The mean, $\bar{x}$, of a set of six numbers is 85.5. Find the value of Σx.
b) One more data value is added to the set. The new mean is 84.9. Find the data value that was added.

Exam Question

Q1 In a supermarket, two types of chocolate drops were compared.
The weights, a grams, of 20 chocolate drops of brand A are summarised by: $\Sigma a = 60.3$ g, $\Sigma a^2 = 219$ g^2
The mean weight of 30 chocolate drops of brand B was 2.95 g, and the standard deviation was 1 g.

a) Find the mean weight of a brand A chocolate drop. [1 mark]

b) Find the standard deviation of the weight of the brand A chocolate drops. [2 marks]

c) Compare the weights of chocolate drops from brands A and B. [2 marks]

d) Find the standard deviation of the weight of all 50 chocolate drops. [4 marks]

People who enjoy this stuff are standard deviants...

You don't always need to find values like Σx^2 or Σfx — sometimes they're given in exam questions. Your calculator can probably find the mean and standard deviation of a data set too. But you also need to understand the formulas, so make sure you can do these calculations by hand — you could easily get asked to show your working.

Grouped Data

*Sometimes, some helpful person will put a set of data into **groups** for you. Which isn't always actually very helpful...*

To draw a Histogram find the Frequency Density

Histograms can be drawn for **continuous** data that is **grouped** into 'classes'. As you saw on page 145, you need to plot the **frequency density**, which is found using: frequency density = **frequency** ÷ **class width**.

Example: Draw a histogram to represent the data in this table, showing the masses of parcels.

Mass (to nearest 100 g)	100-200	300-400	500-700	800-1100
Number of parcels	100	250	600	50

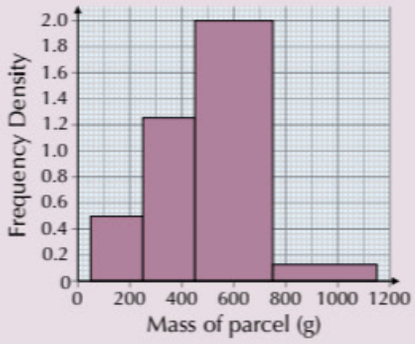
First draw a table showing the **upper and lower class boundaries**, plus the **frequency density**:

Smallest mass of parcel that will go **in that class**. **Biggest** mass that will go **in that class**.

Mass of parcel (to nearest 100 g)	Lower class boundary (lcb)	Upper class boundary (ucb)	Class width = ucb − lcb	Frequency	Freq. density = frequency ÷ class width
100-200	50	250	200	100	0.5
300-400	250	450	200	250	1.25
500-700	450	750	300	600	2
800-1100	750	1150	400	50	0.125

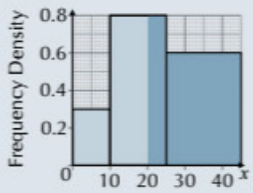
Look — no gaps between each ucb and the next lcb.

Now you can draw the histogram:



You can use **histograms** or grouped frequency tables to **estimate** the number of readings in a **given range**.

Example: Estimate the number of readings above $x = 20$ on the histogram on the right.

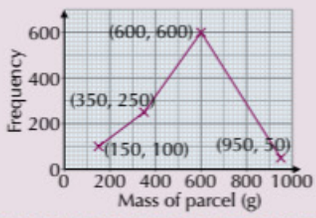


The **area** of the bar is the **total frequency** for that class. You're interested in the **last third** of the class 10–25, and the whole of the class 25–45, so add those two areas:

$(15 \times 0.8) \div 3 + 20 \times 0.6 = \mathbf{16}$

You can also show grouped data on a **frequency polygon**. For each class, plot the **class midpoint** against the **frequency**, then join the points with **straight lines**.

This is a frequency polygon for the parcels data above:



For some people, studying histograms is always a classy affair.

If the data's Grouped you can Estimate the Mean

For **grouped data**, you can't find the mean, median or mode **exactly**. You have to estimate them instead.

Example: The data in this table represents the heights of a number of trees. Estimate the mean of these heights.

Height (to nearest m)	0-5	6-10	11-15	16-20
Number of trees	26	17	11	6

Here, you assume that every reading in a class takes the **class midpoint** (which you find by adding the **lower class boundary** to the **upper class boundary** and **dividing by 2**). It's best to make another table...

Height (to nearest m)	Class midpoint (x)	Number of trees (f)	fx
0-5	2.75	26 (26)	71.5
6-10	8	17 (43)	136
11-15	13	11	143
16-20	18	6	108
	Totals	60 (= Σf)	458.5 (= Σfx)

Lower class boundary = 0
Upper class boundary = 5.5
So the mid-class value
= (0 + 5.5) ÷ 2 = 2.75

Estimated mean $= \frac{458.5}{60}$
$= \mathbf{7.64\ m}$ (3 s.f.)

I've added running totals here — you don't need them for this question, but trust me, they'll come in handy over the next couple of pages...

With grouped data you can't find the mode — only the **modal class**. If all the classes are the **same width**, this is the class with the **highest frequency**. If the classes have different widths, it's the class with the **highest frequency density**. In this example, the modal class is **0-5 m**.

Grouped Data

To *Estimate* the *Median* of Grouped Data, use *Linear Interpolation*

Linear interpolation works by assuming the readings in each class are **evenly spread**.
Here's how you use it to estimate the **median**:

1) Find $n \div 2$ (the position of the median), and work out which class the median falls in.
2) For the class the median is in, the estimated median is:

$$\text{lcb of class} + \text{width of class} \times \frac{\text{position of median} - \text{number of frequencies before class}}{\text{number of readings in class}}$$

You can just use $n \div 2$ here for any n — don't worry about the rules on p.142 in this case.

Example: Using the data from the previous example, estimate the median height of the trees.

$n \div 2 = 60 \div 2 = 30$, so the 'running total' tells you the median must be in the '**6-10**' class.
For the class 6-10, the lcb is 5.5 and the ucb is 10.5 so the width of the class is $10.5 - 5.5 = 5$.

position of median = $n \div 2$ — 26 values before the class 6–10

So **estimated median** $= 5.5 + 5 \times \frac{30-26}{17} = 5.5 + 5 \times \frac{4}{17} =$ **6.68 m** (3 s.f.)

This fraction is the proportion of the way through the class that you'd expect to find the median.

You can estimate the *Standard Deviation* too

Like for the mean, you can estimate the **variance** and **standard deviation** by assuming every reading takes the value of the **class midpoint** and using the **frequencies** to estimate Σfx and Σfx^2.

Example: Estimate the standard deviation of the heights of the trees in the table on the previous page.

You need to add two more columns to the table you had for the mean, for x^2 and fx^2:

Height (to nearest m)		x^2	fx^2
0-5		7.5625	196.625
6-10		64	1088
11-15		169	1859
16-20		324	1944
			5087.625 (= Σfx^2)

Now you've got the totals in the table, you can calculate estimates for the variance and the standard deviation:

variance $\approx \frac{\Sigma fx^2}{\Sigma f} - \overline{x}^2 = \frac{5087.625}{60} - (7.64...)^2$
$= 26.39...$ m²

so **standard deviation** $\approx \sqrt{26.39...}$
$=$ **5.14 m** (3 s.f.)

Use the unrounded values for the estimated mean and variance here.

Practice Questions

Q1 The table on the right shows the lengths in minutes (to the nearest minute) of twenty phone calls. Draw a histogram of the data.

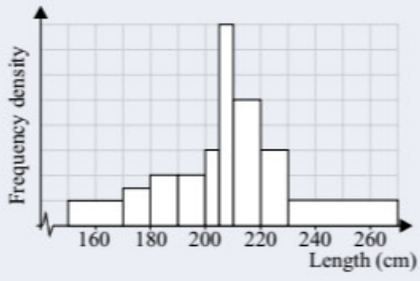
Call length (mins)	0-2	3-5	6-8	9-15
Number of calls	10	6	3	1

Q2 The speeds of 60 cars travelling in a 40 mph speed limit area were measured to the nearest mph. The data is summarised in the table. Calculate estimates of the mean and median, and state the modal class.

Speed (mph)	30-34	35-39	40-44	45-50
Frequency	12	37	9	2

Exam Questions

Q1 The histogram on the right shows the nose-to-tail lengths of 50 lions in a nature reserve. Find the number of lions measuring over 220 cm. [3 marks]



Q2 The profits of 100 businesses are given in this table.

Profit (£p million)	$4.5 \le p < 5.0$	$5.0 \le p < 5.5$	$5.5 \le p < 6.0$	$6.0 \le p < 6.5$	$6.5 \le p < 8.0$
No. of businesses	21	26	24	19	10

a) Represent the data in a histogram. [3 marks]
b) Estimate the mean and standard deviation of this data. [5 marks]
c) Use linear interpolation to estimate the median profit. [3 marks]

My pop group 'Data Days' topped the charts and won 8 histogrammys...

A class with lower class boundary = 50 g and upper class boundary = 250 g can be written in different ways. You might see: "100-200 to nearest 100 g", or "50 ≤ mass < 250", or "50-" followed by "250-" for the next class, etc. They all mean the same — make sure you know how to spot the lcb and ucb for each version.

All you standard deviation lovers out there will be happy to hear there are even more measures of variation to learn.

The **Range** is a Measure of **Variation**...

The **range** is about the simplest measure of variation you could imagine: **Range** = highest value – lowest value

But the range is heavily affected by **extreme values**, so it isn't really the most useful way to measure variation.

Quartiles divide the data into **Four**

You've seen how the **median** divides a data set into **two halves**. Well, the **quartiles** divide the data into **four quarters** — with 25% of the data less than the **lower quartile**, and 75% of the data less than the **upper quartile**.
There are various ways you can find the **quartiles**, but if you use the method below, you'll be fine.

1 **To find the lower quartile (Q_1), work out $\frac{n}{4}$.**

(i) If $\frac{n}{4}$ is a **whole number**, the lower quartile is the **average of this term and the one above**.

(ii) If $\frac{n}{4}$ is **not a whole number**, **round the number up** to find the position of the lower quartile.

2 **To find the upper quartile (Q_3), work out $\frac{3n}{4}$.**

(i) If $\frac{3n}{4}$ is a **whole number**, the upper quartile is the **average of this term and the one above**.

(ii) If $\frac{3n}{4}$ is **not a whole number**, **round the number up** to find the position of the upper quartile.

Example: Find the median and quartiles of the following data:
2, 5, 3, 11, 6, 8, 3, 8, 1, 6, 2, 23, 9, 11, 18, 19, 22, 7

The median is also known as Q_2.

First put the list **in order**: 1, 2, 2, 3, 3, 5, 6, 6, 7, 8, 8, 9, 11, 11, 18, 19, 22, 23

You need to find Q_1, Q_2 and Q_3, so work out $\frac{n}{4} = \frac{18}{4}$, $\frac{n}{2} = \frac{18}{2}$, and $\frac{3n}{4} = \frac{54}{4}$.

1) $\frac{n}{4}$ is **not** a whole number (= 4.5), so round up and take the 5th term: $Q_1 = \mathbf{3}$

2) $\frac{n}{2}$ **is** a whole number (= 9), so find the average of the 9th and 10th terms: $Q_2 = \frac{7+8}{2} = \mathbf{7.5}$

3) $\frac{3n}{4}$ is **not** a whole number (= 13.5), so round up and take the 14th term: $Q_3 = \mathbf{11}$

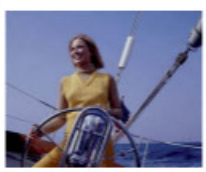
The **Interquartile Range** is another measure of **Variation**

Interquartile range (IQR) = upper quartile (Q_3) – lower quartile (Q_1)

The IQR shows the range of the 'middle 50%' of the data.

Example: Find the interquartile range of the data in the previous example.

$Q_1 = 3$ and $Q_3 = 11$, so the interquartile range = $Q_3 - Q_1 = 11 - 3 = \mathbf{8}$



Next stop: the Percent Isles.

Percentiles divide the data into **100**

Percentiles divide the data into 100 — the median is the **50th percentile**, Q_1 is the **25th percentile**, etc.

For example, the **position** of the 11th percentile (P_{11}) is $\frac{11}{100}$ × total frequency.

You find **interpercentile ranges** by **subtracting** two percentiles, e.g. the 20% to 80% interpercentile range = $P_{80} - P_{20}$.

If your data is **grouped**, you might need to use **linear interpolation** to find the quartiles or percentiles.
The method's the same as the one on p.147, just swap 'median' for the percentile or quartile you want.

Example: Estimate the 80th percentile for the tree data at the bottom of p.146.

$\Sigma f = 60$, so the position of the 80th percentile is $\frac{80}{100} \times 60 = 48$, so P_{80} is in the '**11-15 m**' class.

Using the linear interpolation formula: $P_{80} = 10.5 + 5 \times \frac{48 - 43}{11} \Rightarrow P_{80} = \mathbf{12.8\ m}$ (3 s.f.)

(l.c.b. = 10.5; class width = 5; frequency before class = 43; frequency in class = 11)