9. Statistics

YOUR NOTES ⬇

CONTENTS

9.1 DISPLAYING DATA

9.1.1 STEM & LEAF DIAGRAMS

9. Statistics

YOUR NOTES ⬇

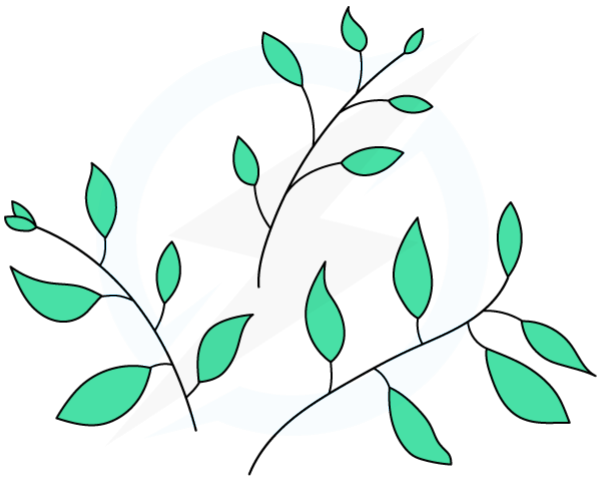
What is a stem-and-leaf diagram?

- A **stem–and–leaf diagram** is a simple but effective way of showing data
- It puts the data into **order**, puts it into **classes** (**groups**) and we can quickly see patterns
- As the data is in order it is also useful for finding the **median** and **quartiles**

What do I need to know?

- Stem-and-leaf diagrams are particularly useful for two-digit data but can be used for bigger numbers
- Two-digit data could be something like 26 but could also be 2.6, due to this one of the essential things about a stem-and-leaf diagram is that it has a key
- You may also come across back-to-back stem-and-leaf diagrams which are used to compare two sets of data

1. Stem-and-leaf diagrams



9. Statistics

YOUR NOTES ⬇

- The digits from the data are split into two – stems and leaves
- As in nature though, a stem can have more than one leaf, so the stems become our classes in our data
- Eg The data value 26 would be split into a stem of 2 and a leaf of 6
  That will then mean the "2" becomes a class interval – ie the 20's
  Any other values in the 20's would join the same class – so a stem of 2 would have two leaves
  Eg. Draw a stem-and-leaf diagram for the following data

26 45 32 27 29 30 40 36 37

As the data is not in order draw a **rough** diagram first to get the data values into the correct format:

Stem	Leaves			
2	6	7	9	
4	5	0		
3	2	0	6	7

Now put the stems and their leaves in order

Stem	Leaves			
2	6	7	9	
3	0	2	6	7
4	0	5		

Key: 2|6 means 26

Add a key so we know what the data is showing

9. Statistics

YOUR NOTES ⬇

Worked Example

1. A hospital is trying to compare two different medications that claim to reduce blood pressure. They give one set of patients "Drug 1" and a second set of patients "Drug 2" and three hours later record the amount the blood pressure of every patient is reduced by. The results for both groups are below.

Drug 1

12 31 24 18 21 34 40 19 23 17 16

Drug 2

24 18 29 27 32 36 34 31 28 31

(a) Draw a back-to-back stem-and-leaf diagram to show these results,

(b) Comment briefly on what drug you think is more effective, giving a reason why.

9. Statistics

YOUR NOTES ⬇

(a)

Drug 2						Drug 1					
				8	1	2	8	9	7	6	
1	1	4	6	2	3	1	4				
	8	7	9	4	2	4	1	3			
					4	0					

Rough:

Notice how the Drug 2 leaves 'grow' from the centre outwards so when you do an ordered diagram the lowest values will be closest to the stems

Final:

Drug 2						**Drug 1**					
				8	**1**	**2**	**6**	**7**	**8**	**9**	
	9	**8**	**7**	**4**	**2**	**1**	**3**	**4**			
6	**4**	**2**	**1**	**1**	**3**	**1**	**4**				
					4	**0**					

Key: 4|2 means a blood pressure reduction of 42

Take your time to make sure you have all the leaves and that they are in order with the correct leaves (also in order)

Don't forget the key!

(b)

Drug 2 is more effective at reducing blood pressure as it has a median reduction of 30 whereas Drug 1 has a median reduction of 21.

There are a few options here but it is important you give a reason to justify your decision

You could say Drug 2 as it had most values in the 30's whereas drug 1 had most values in the 10's

This option would show you understand how a stem and leaf diagram splits data into classes/groups

9. Statistics

YOUR NOTES ⬇

9.1.2 FREQUENCY POLYGONS

What are frequency polygons?

- **Frequency polygons** are a very simple way of showing frequencies for **continuous**, **grouped** data and give a quick guide to how frequencies change from one class to the next

What do I need to know?

- Apart from plotting and joining up points with straight lines there are 2 rules for frequency polygons:
  - Plot points at the **MIDPOINT** of class intervals
  - Unless one of the frequencies is 0 do not join the frequency polygon to the x-axis, and do not join the first point to the last one
- The result is not actually a polygon but more of an open one that 'floats' in mid-air!
- You may be asked to draw a frequency polygon and/or use it to make comments and compare data

1. Drawing

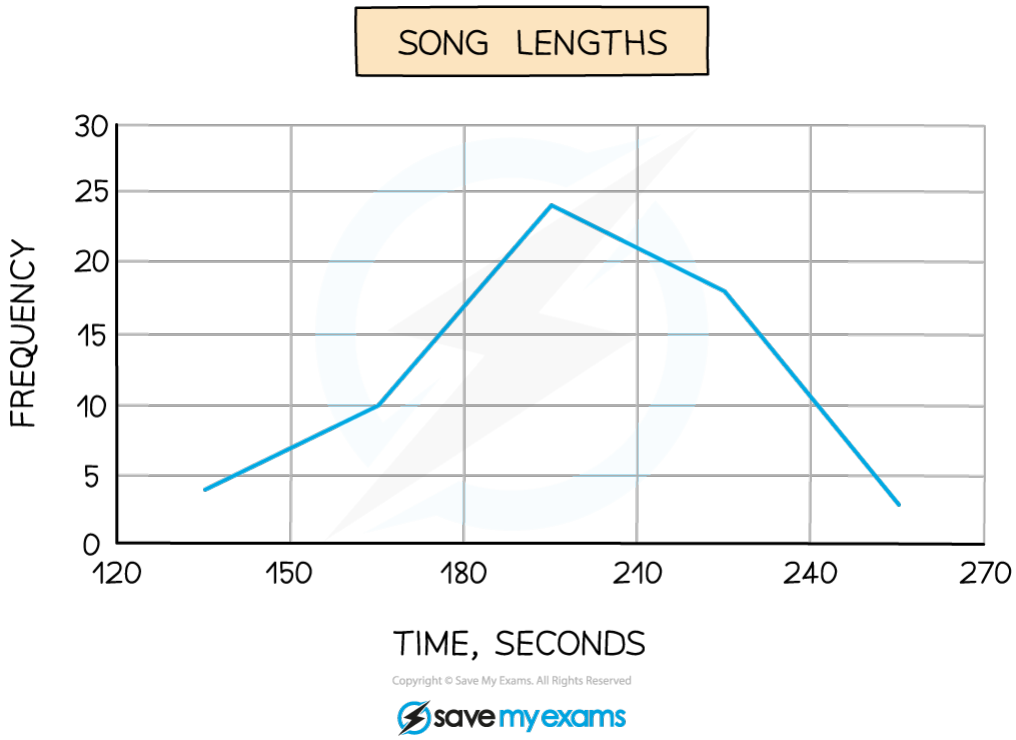
- The lengths of 60 songs, in seconds, are recorded in the table below

Song length, t seconds	Frequency
$120 \leq t < 150$	4
$150 \leq t < 180$	10
$180 \leq t < 210$	24
$210 \leq t < 240$	18
$240 \leq t < 270$	3

9. Statistics

YOUR NOTES

Draw a frequency polygon for these data:



2. Using and interpreting

- What can you say about the data above, particularly by looking at the diagram only?
  - The two things to look for are **averages** and **spread**
  - The **modal class** is $180 \leq t < 210$
  - It would be acceptable to say that 195 seconds is the **modal** song length
  - The diagram (rather than the table) shows the **range** of song lengths is 255 – 135 = 120 seconds
  - If 2 frequency polygons are drawn on the same graph comparisons between the 2 sets of data can be made

Exam Tip

Jot down the midpoints next to the frequencies so you are not trying to work them out in your head while also concentrating on actually plotting the points.

9. Statistics

YOUR NOTES ⬇

Worked Example

1. A local council ran a campaign to encourage households to waste less food.

To compare the impact of the campaign the council recorded the weight of food waste produced by 30 households in a week both before and after the campaign.

The results are shown in the table below.

Food Waste w kg	Frequency Before Campaign	Frequency After Campaign
$1 \leq w < 1.4$	3	5
$1.4 \leq w < 1.8$	5	8
$1.8 \leq w < 2.2$	8	14
$2.2 \leq w < 2.6$	12	3
$2.6 \leq w < 3$	2	1

(a) On the same diagram, draw two frequency polygons, one for before the council's campaign and one for after.

(b) Comment on whether you think the council's campaign has been successful or not and give a reason why.

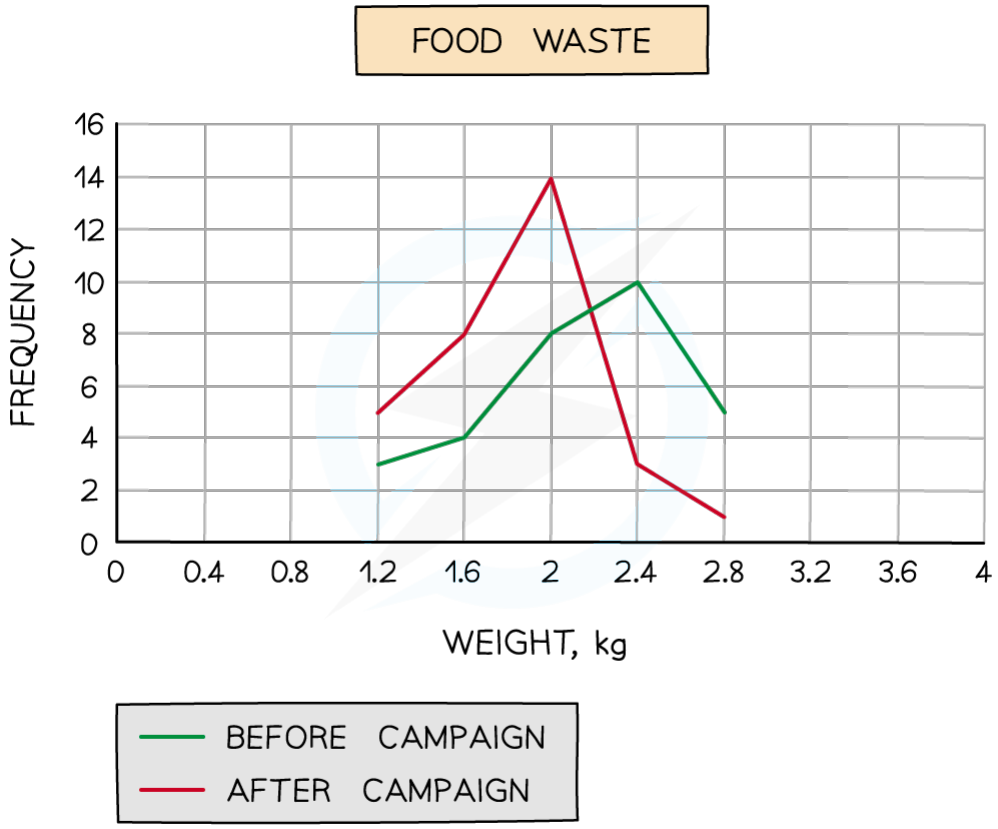
(a)

Midpoints are: 1.2, 1.6, 2, 2.4 and 2.8

1 - Jot down the midpoints so you can focus on plotting points!

9. Statistics

YOUR NOTES ⬇



Remember a key to show which frequency polygon is which

(a)

The council campaign has been successful as the mode amount of waste has reduced from 2.4 kg of food waste per week to 2 kg.

2 - Remember to look for averages and/or spread – in this case the range is the same

Any comment justified by the diagram/maths would be correct

9. Statistics

YOUR NOTES ⬇

9.1.3 SCATTER GRAPHS

What is a scatter graph all about?

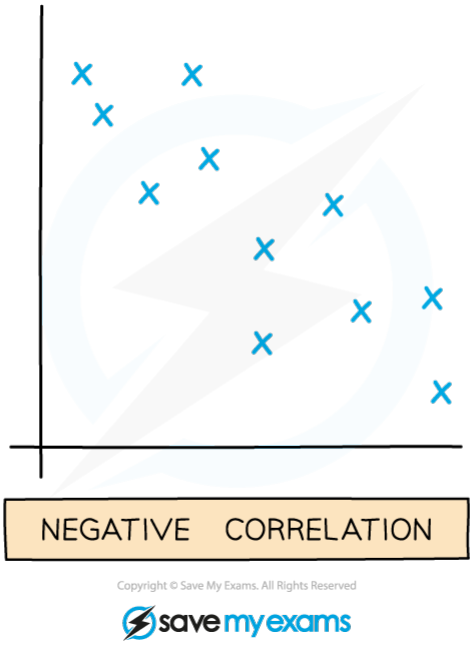
- **Scatter graphs** are used to see if there is a connection between two pieces of data
- For example a teacher may want to see if there is a link between grades in mathematics tests and grades in physics tests
- If there is a connection we can use a **line of best fit** to predict one data value (say the physics grade for example) from a known data value (maths grade)

What do I need to know?

- You need to be able to plot and interpret a scatter graph
- They are sometimes called scatter plots or scatter diagrams but these all mean the same thing
- You also need to know about correlation (positive, negative) and how to draw and use a line of best fit

9. Statistics

YOUR NOTES



9. Statistics

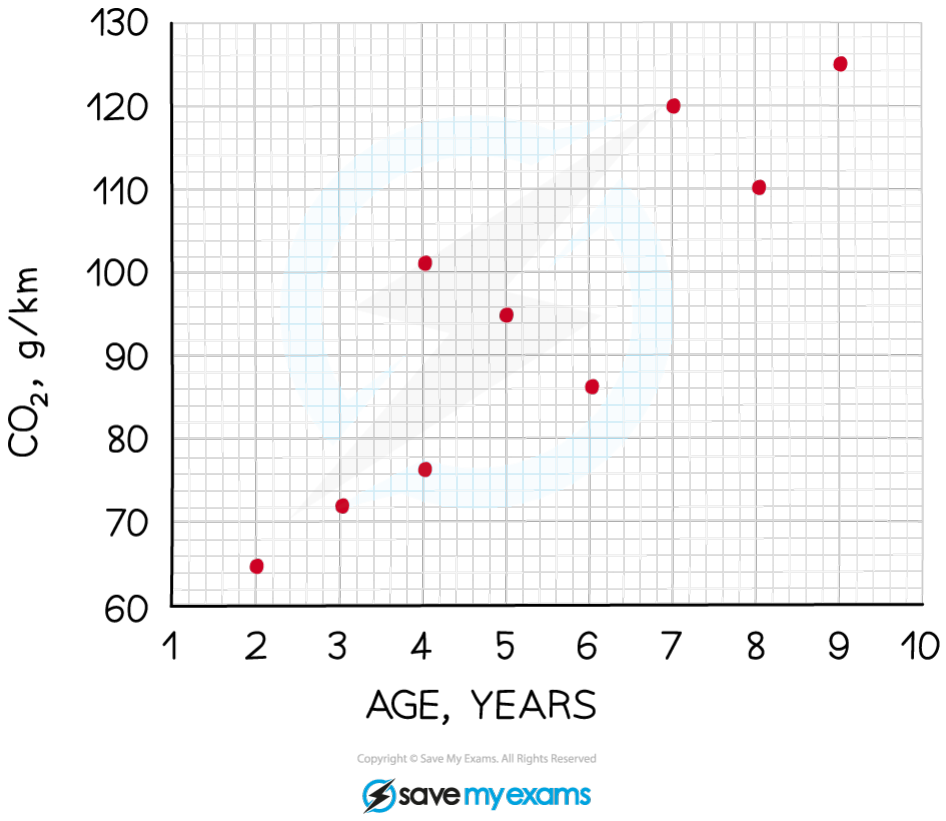
YOUR NOTES ⬇

1. Drawing a scatter graph

- This is simply a matter of plotting points but be very careful which way round you are plotting them particularly when the values are very similar

Age (years)	2	7	4	9	5	6	4	8	3
CO_2 (g/km)	65	120	100	125	94	86	76	110	72

- For example, John is buying a second-hand car but is concerned about carbon dioxide (CO_2) emissions
- He collects data about the age of a car the amount of CO_2 the car emits
- Plot John's results on a scatter graph and decide if there is a connection between age and CO_2 emissions
- If there is use your graph to estimate how old a car emitting 104g/km would be

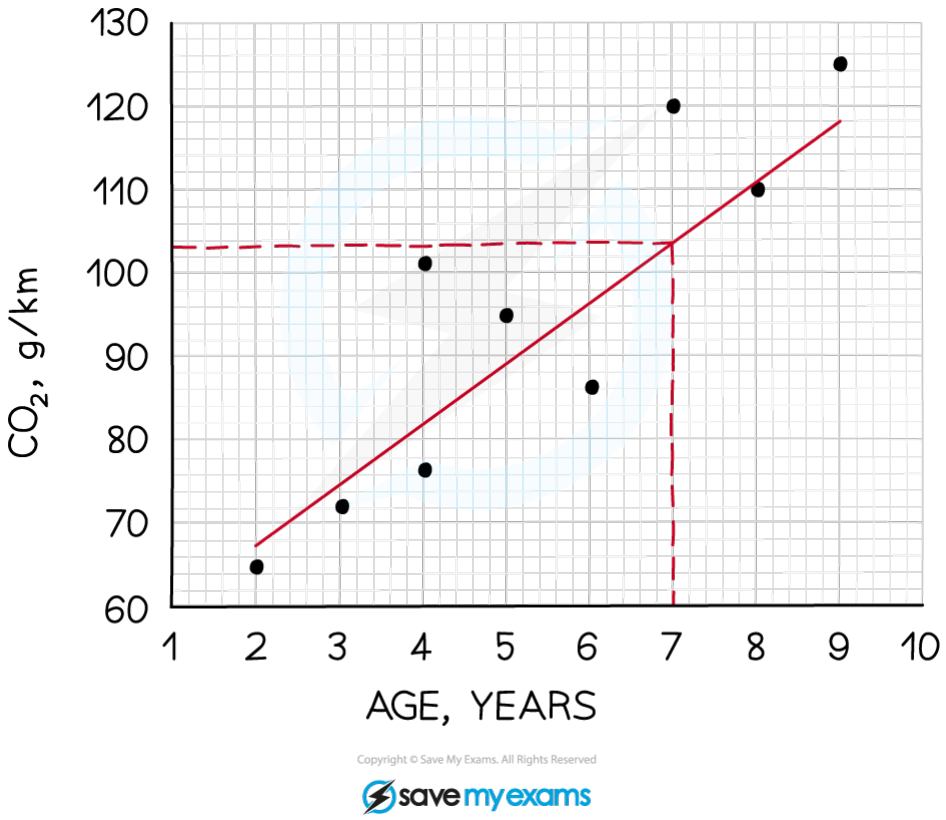


9. Statistics

YOUR NOTES

2. Using a scatter graph

- There is a **positive correlation** (as one value increases, so does the other) so add in a line of best fit and use this to answer the question



- Notice that the line of best fit does not have to go through any of the data points
- From the line of best fit that a car with emissions of 104 g/km would be about 7 years old

Exam Tip

Watch out for **outliers** on scatter graphs – these are rogue results or values that do not follow the general pattern of the data. You should not consider these points when judging where to draw your line of best fit. Ignore it for this purpose. You'd usually only see one of these in a question, if any.

9. Statistics

YOUR NOTES ⬇

Worked Example

1. Sophie is investigating the price of computers to see if the more they cost, the quicker they are. She tests 8 computers and runs the same program on each, measuring how many seconds each takes to complete the program. Sophie's results are shown in the table below.

Price (£)	320	300	400	650	250	380	900	700
Time (seconds)	3.2	5.4	4.1	2.8	5.1	4.3	2.6	3.7

(a) Draw a scatter graph to show this information,

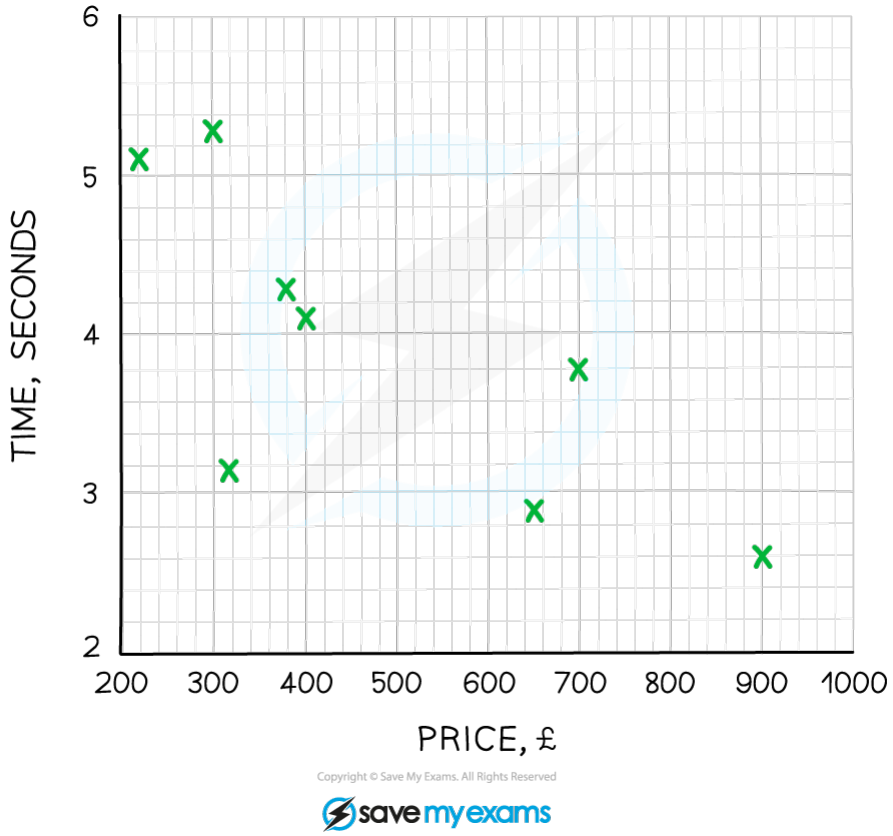
(b) Describe the correlation and explain what this means in terms of the question,

(c) Showing your method clearly estimate the price of a computer that completes the task in 3.5 seconds.

(a)

9. Statistics

YOUR NOTES ⬇



1 - Plot the points carefully and accurately as to not miss any out

(b)

The graph shows negative correlation.

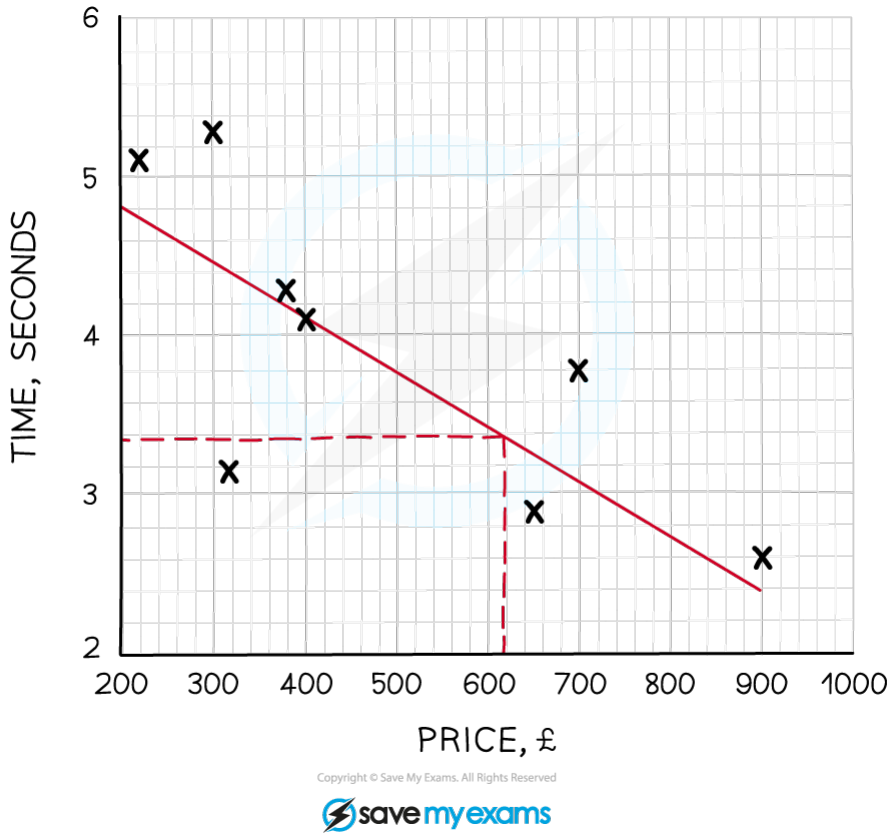
This means that the more a computer costs, the quicker it is at running the program.

2 - As you were asked to explain what the correlation means in terms of the question you need to mention the connection between cost and speed

9. Statistics

YOUR NOTES

(c)



The price of a computer taking 3.5 seconds to run the program should cost around £612.

2 - Depending on the scale/value/etc you may not be able to take an exact reading from your graph, this is fine as a range of answers will be acceptable

9. Statistics

YOUR NOTES ⬇

9.1.4 HISTOGRAMS

Aren't histograms just really hard bar charts?

- No! There are many mathematical differences that you should be aware of but the key difference between a bar chart and a **histogram** is that with a histogram it is the **area of the bars** that determine the frequency, on a bar chart it's the height (or length) of them
- Digging deeper, histograms are used for **continuous** data (bar charts for discrete) and are particularly useful when data has been **grouped** into **different** sized **classes**
- But the key thing to getting started is that it is the area of the bars that tell us what is happening with the data
- This means, unlike any other graph or chart you have come across, it is very difficult to tell anything from simply looking at a histogram, you have to drill down into the numbers and detail

What do I need to know?

- You need to know how to **draw** a histogram (most questions will get you to finish an incomplete histogram)
- When drawing histograms we will need to use **frequency density** (fd):

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

- You'll also need to be able to work backwards from a given histogram to find frequencies and **estimate** the mean

1. Drawing a histogram

- From a given table you need to work out the frequency density for each class
- Then you can plot the data against frequency density with frequency density on the -axis
- For example, plot a histogram for the following data regarding the average speed travelled by trains

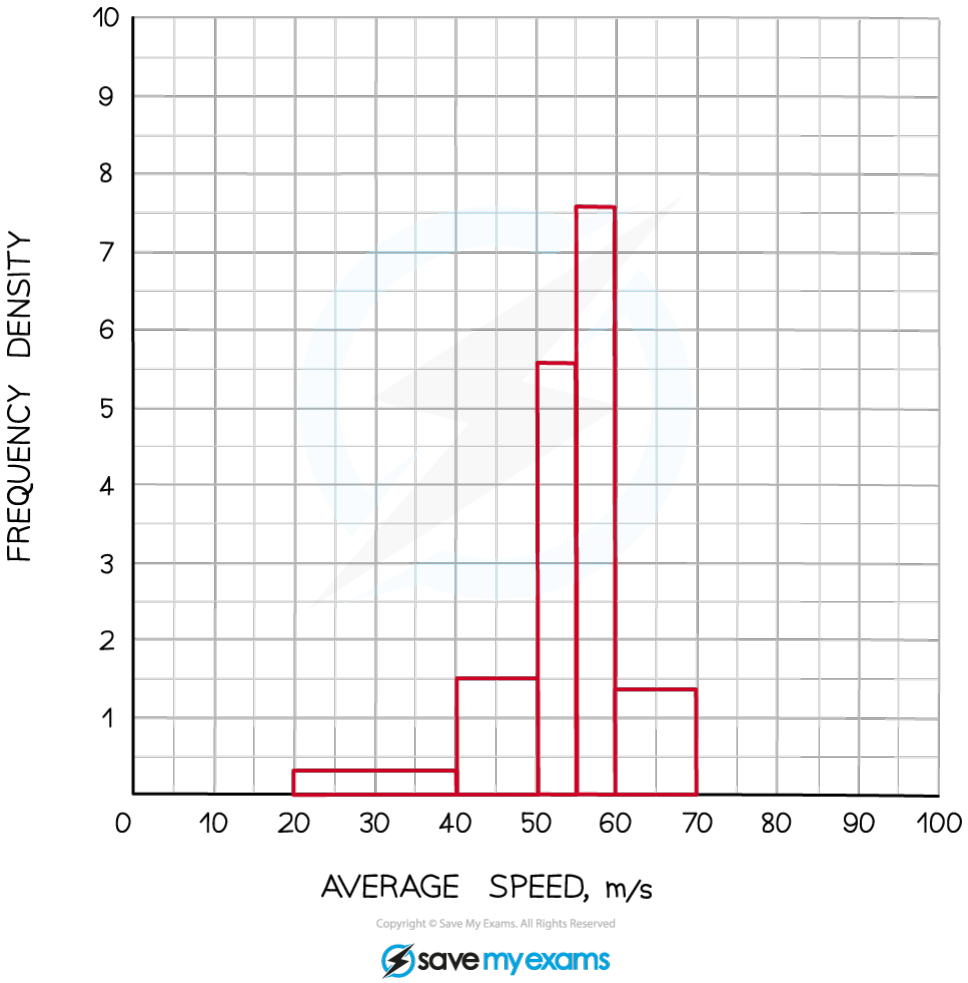
9. Statistics

YOUR NOTES ⬇

Average speed, s m/s	Frequency	Class width	Frequency Density
$20 \leq s < 40$	5	40 – 20 = 20	5 ÷ 20 = 0.25
$40 \leq s < 50$	15	50 – 40 = 10	15 ÷ 10 = 1.5
$50 \leq s < 55$	28	55 – 50 = 5	28 ÷ 5 = 5.6
$55 \leq s < 60$	38	60 – 55 = 5	38 ÷ 5 = 7.6
$60 \leq s < 70$	14	70 – 60 = 10	14 ÷ 10 = 1.4

- Note that the class width column isn't essential but it is crucial you show the frequency densities

- Now we draw bars (touching, as the data (speed) is continuous) with widths of the class intervals and heights of the frequency densities



9. Statistics

YOUR NOTES ⬇

2. Interpreting histograms

- We shall still use the example above here but shall pretend we never had the table of data and were only given the finished histogram
- To **estimate** the **mean**:
  - You need to know the total frequency and what all the data values add up to
  - You can't find the exact total of the data values as this is grouped data but we can estimate it using **midpoints**

Since:

$$\textit{frequency density} = \frac{frequency}{class\ width}$$

then it is easy to rearrange to see that:

frequency = frequency density × class width

$20 \leq s < 40$:	frequency = 0.25 x 20 = 5	midpoint = 30
$40 \leq s < 50$:	frequency = 1.5 x 10 = 15	midpoint = 45
$50 \leq s < 55$:	frequency = 5.6 x 5 = 28	midpoint = 52.5
$55 \leq s < 60$:	frequency = 7.6 x 5 = 38	midpoint = 57.5
$60 \leq s < 70$:	frequency = 1.4 x 10 = 14	midpoint = 65

Total of frequencies = 5 + 15 + 28 + 38 + 14 = 100

- You can draw all of the above in a table if you wish
- Now you can total up (an estimate of) the data values and find the mean:

Total = 5 x 30 + 15 x 45 + 28 x 52.5 + 38 x 57.5 + 14 x 65 = 5390

(Be careful if you type all this into your calculator in one go!)

Estimate of Mean = 5390 ÷ 100 = 53.9

9. Statistics

YOUR NOTES ⬇

Exam Tip

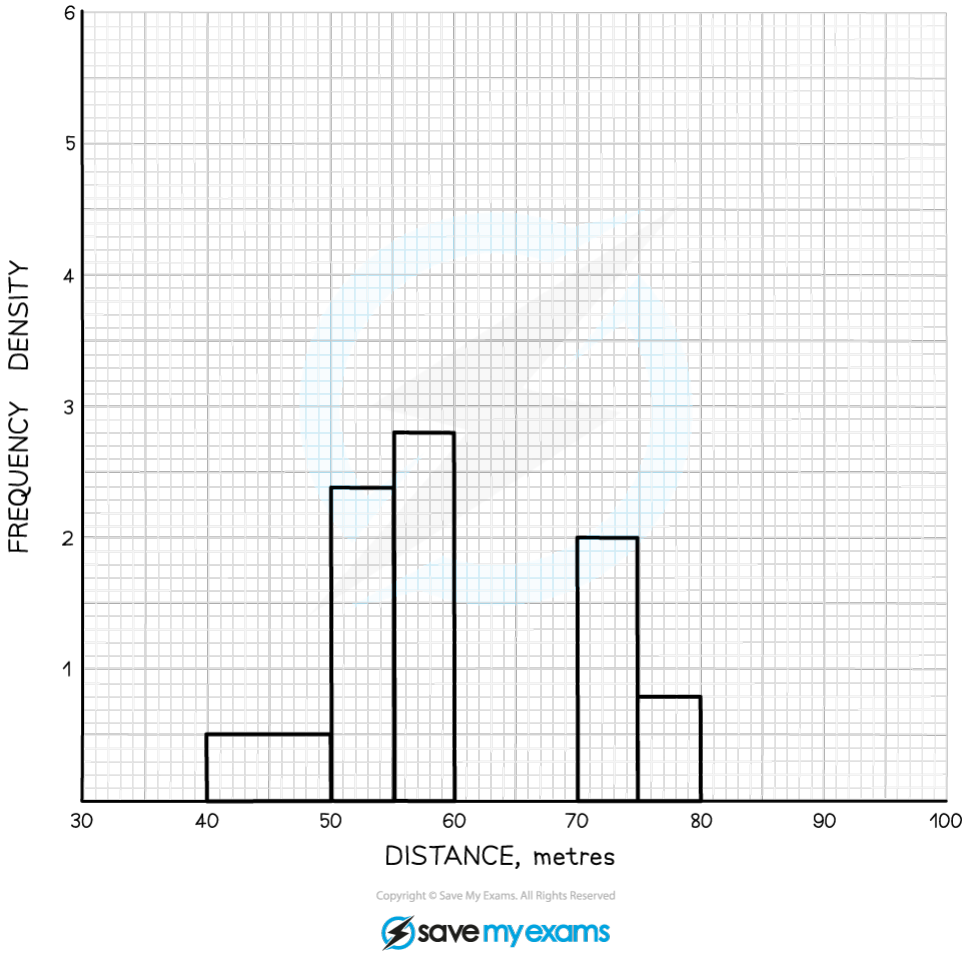
Always work out and write down the frequency densities. Many students lose marks in exams as they go straight to the graph when asked to draw a histogram and they mess up the calculations. Method marks are available for showing you know to use frequency density rather than frequency.

Worked Example

1. A histogram is shown below representing the distances achieved by some athletes throwing a javelin.

9. Statistics

YOUR NOTES ⬇



(a) There are two classes missing from the histogram. These are:

Distance, x m	Frequency
$60 \leq x < 70$	8
$80 \leq x < 100$	2

Add these to the histogram

(b) Approximately how many athletes threw the javelin a distance over 75 metres.

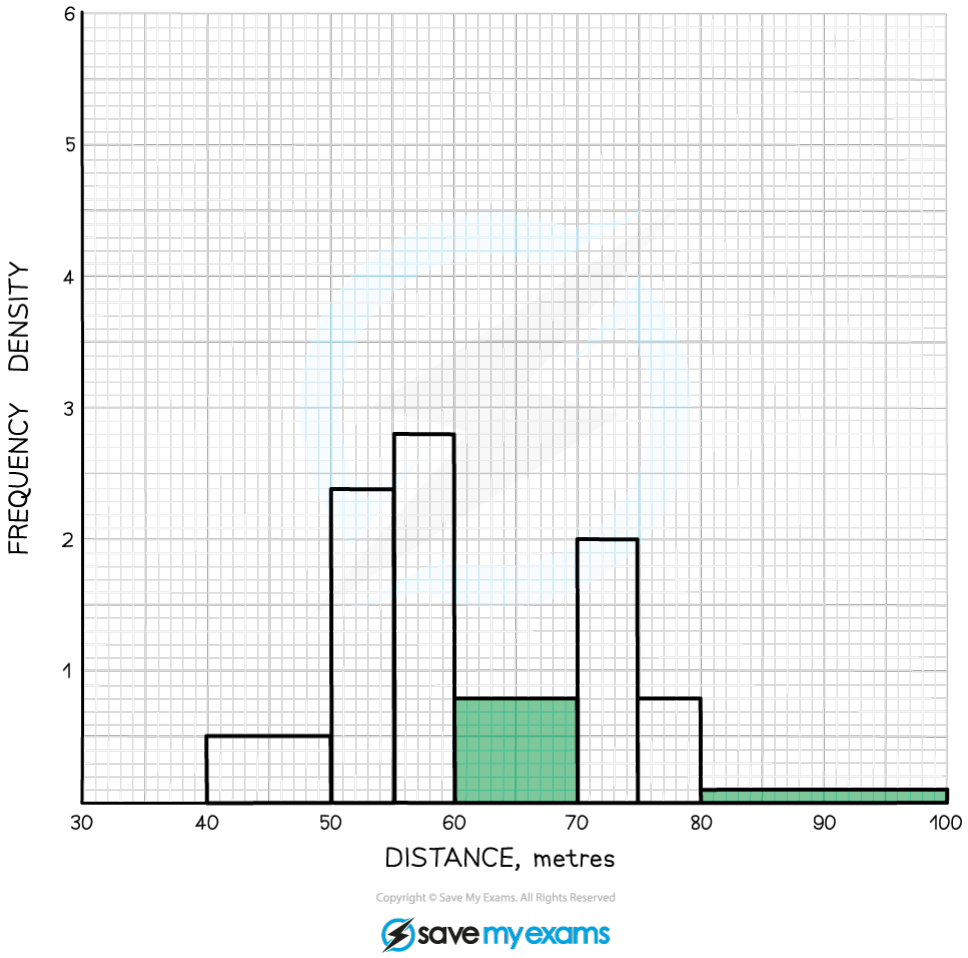
Distance, x m	Frequency	Frequency Density
$60 \leq x < 70$	8	8 ÷ 10 = 0.8
$80 \leq x < 100$	2	2 ÷ 20 = 0.1

9. Statistics

YOUR NOTES

(a)

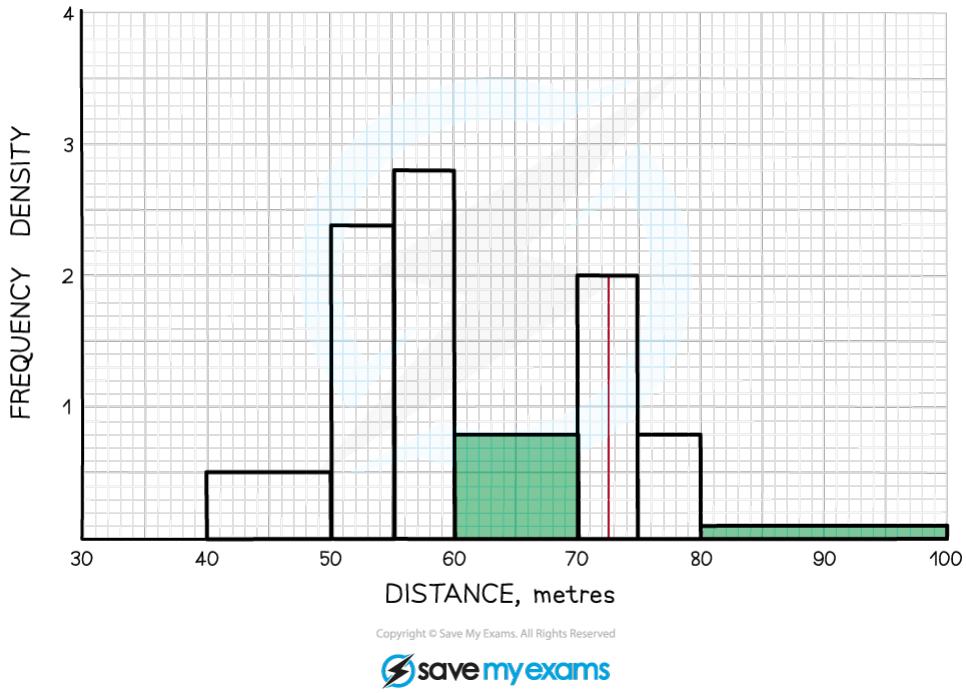
1 - Remember to clearly show you've worked out the missing frequency densities



9. Statistics

YOUR NOTES

(b)



2 - You need to work out the frequency of the bars that are at 75m or greater

$70 \leq x < 75$: frequency = 5 x 2 = 10

$75 \leq x < 80$: frequency = 5 x 0.8 = 4

However 75 falls in the middle of one of the groups – so you would take half of that frequency

½ x 10 + 4 + 2 = 11

We know already that the frequency for $80 \leq x < 100$ is 2

Number of athletes throwing over 75 m is 11

9. Statistics

YOUR NOTES ⬇

9.2 MEAN/MEDIAN/MODE/RANGE

9.2.1 MEAN, MEDIAN & MODE

Why do we have different types of average?

- You'll hear the phrase "on average" used a lot, from politicians talking about the economy to sports analysts to shops talking about their "average customer"
- However not all data is numerical (eg the party people voted for in the last election) and even when it is numerical, some of the data may lead to misleading results
- This is why we have **3 types of average**

What do I need to know?

1. Mean

- This is what is usually meant by "average" – it's like an ideal world where everybody has the same, everything is shared out equally
- It is the **TOTAL** of all the values **DIVIDED** by the **NUMBER OF VALUES**
  Find the mean of 4, 6, 7, 9

  4 + 6 + 7 + 9 = 26

  26 ÷ 4 = 6.5

  Mean = 6.5
- Problems with the mean occur when there are one or two unusually high (or low) values in the data (**outliers**) which can make the mean too high (or too low) to reflect any patterns in the data

9. Statistics

YOUR NOTES ⬇

2. Median

- This is similar to the word medium, which can mean in the middle
- So the median is the middle value – but beware, the data has to be arranged into numerical order first
  Find the median of 20, 43, 56, 78, 92, 56, 48

  In order: 20, 43, 48, 56, 56, 78, 92

  To find the median cross out numbers from either end until you meet in the middle (cross them out lightly so you can still read them)
  This may not be necessary with small lists but is more important when working with lots of data

  ~~20~~, 43, 48, 56, 56, 78, ~~92~~

  ~~20~~, ~~43~~, 48, 56, 56, ~~78~~, ~~92~~

  ~~20~~, ~~43~~, ~~48~~, 56, ~~56~~, ~~78~~, ~~92~~

  Median = 56
- We would use the median instead of the mean if we did not want extreme values (outliers) affecting our data
- If we have an even number of values we would get two values in the middle
- In these cases we take the half-way point between these two values. This is usually obvious but, if not, we **add the two middle values and divide by 2** (this is the same as finding the mean of the middle two values)
  20, 43, 46, 48, 56, 56, 78, 92 (as above with an extra 46 in there!)

  When crossed out we get

  ~~20~~, ~~43~~, ~~46~~, 48, 56, ~~56~~, ~~78~~, ~~92~~

  So the two middle values are 48 and 56.

  Halfway is 52 but if you can't spot that you can work it out …

  48 + 56 = 104

  104 ÷ 2 =52

  Median = 52

9. Statistics

YOUR NOTES ⬇

3. Mode

- Not all data is numerical and that is where we use mode
- **MO**de means the **M**ost **O**ften
- So it is often used for things like "favourite ..." or "... sold the most" or "... were the most popular"
- Mode is sometimes referred to as **modal** – so you may see phrases like "**modal value**" – but they still mean the mode
  12 people were asked about their favourite crisp flavour. The responses are below:

  Salt and vinegar, Prawn cocktail, Ready salted,

  Ready salted, Salt and vinegar, Salt and vinegar,

  Smokey bacon, Ready salted, Salt and vinegar

  Salt and vinegar, Cheese and onion, Ready salted

  With only a few pieces of data it is quite quick and easy to see here that Salt and vinegar is chosen the most
  With more data it may be wise to create a tally chart or similar to help count the number of each flavour
  Mode is Salt and vinegar
- Be aware that the mode can apply to numerical data as well (from the data used in the example for the median the mode would have been 56)
- Sometimes if no value/data occurs more often than others we say there is no mode
- If two values occur the most we may say there are two modes (**bi-modal**) – whether it is appropriate to do this will depend on what the data is about

Worked Example

1. (a) Briefly explain why the mean is not a suitable average to use in order to analyse the way people voted in the last general election.
   (b) Suggest a better measure of average that can be used.

(a)

Political parties have names and so the data is not numerical

(b)

The mode average can be used for non-numerical data

9. Statistics

YOUR NOTES ⬇

2. 15 students were timed how long it took them to solve a maths problem. Their times, in seconds, are given below.

12	10	15	(37)
14	17	11	(42)
12	13	9	(34)
21	14	20	(55)
19	16	23	(58)

(a) Find the mean and median times.

(b) What can you say about the mode of the data?

(a)

Mean:

12 + 10 + 15 + 14 + 17 + 11 + 12 + 13 + 9 + 21 + 14 + 20 + 19 + 16 + 23 = 226

1 - You could do the adding up in bits by adding the rows (or columns) as above in brackets

226 ÷ 15 = 15.066...

Do the mean in two stages to avoid confusion around using brackets on your calculator and to show all stages of working

Mean = 15.1 (to one decimal place)

Round final answer to something sensible if not asked to

Median:

~~9~~ ~~10~~ ~~11~~ ~~12~~ ~~12~~ ~~13~~ ~~14~~ 14 ~~15~~ ~~16~~ ~~17~~ ~~19~~ ~~20~~ ~~21~~ ~~23~~

Median = 14

2 - Make sure you write them in order and do not miss any out, it's a good idea to lightly cross them off the original list

(b)

Two values occur more than any others, 12 and 14

3 - Notice how this is easy to see once the data is in order from finding the median in part (a)

So we can say there are two modes – 12 and 14, or, we could say that there is no mode

9. Statistics

YOUR NOTES ⬇

9.2.2 AVERAGES FROM TABLES & CHARTS

How do we find averages if there are lots of values?

- In reality there will be far more data to work with than just a few numbers
- In these cases the data is usually organised in such a way to make it easier to follow and understand – for example in a **table** or **chart**
- We can still find the **mean**, **median** and **mode** but have to ensure we understand what the table or chart is telling us

What do I need to know?

- Finding the median and mode from tables/charts is fairly straightforward once you understand what the table/chart is telling you so these notes focus mainly on finding the mean

1. Finding the mean from (discrete) data presented in tables

- Tables allow data to be summarised neatly – and quite importantly it puts it into order

Eg. the number of pets owned by 40 pupils in year 11 are summarised in the table below:

No. of Pets	Frequency
0	12
1	15
2	8
3	3
4	2

Work out the mean and median number of pets per pupil.

The **mean** can be found as you long as understand what the table is telling you:

It tells you:

9. Statistics

YOUR NOTES ⬇

12 (of the 40) pupils had no pets; 15 of them had 1 pet; 8 had 2 pets

This means you can add up all the 0's very quickly, all the 1's very quickly etc. using multiplication: 12 x 0 = 0, 15 x 1 = 15, etc.

The easiest way to do this is to add another column to the table and adding a **total row** will prove useful in the next stage too:

No. of Pets	Frequency	Pets x Frequency
0	12	0 x 12 = 0
1	15	1 x 15 = 15
2	8	2 x 8 = 16
3	3	3 x 3 = 9
4	2	4 x 2 = 8
Total	40	48

- You can now see that the total number of pets for the whole of the 40 pupils is 48
- If you hadn't been told in the question, the total of the frequency column would've told you how many pupils there are
- You can now find the mean:

  48 ÷ 40 = 1.2

  The mean number of pets is 1.2 (pets per pupil)

- You may sometimes see the number of pets called x and the frequency f so the last column would be:
  f x (f times x)

9. Statistics

YOUR NOTES ⬇

2. Median

- The **median** is a little complicated to understand but easy to work out
- Remember you are looking for the middle value when the data is in order
- For the example above the table is in order (0, then 1, 2, etc.) – so you need to work out which of the 40 values is in the median position
- To do this you add one to the number of values and divide by 2

  (40 + 1) ÷ 2 = 20.5 – so the median is in the "20.5th" position

- The table tells you the first twelve numbers on the list are all 0's, the next 15 are all 1's – so the "20.5th" value in the list must be a 1
  The median number of pets is 1

  (Note that the "20.5th" value is referring to halfway between the 20th and 21st values. Both of these are 1's so median must be 1 (or use (1 + 1) ÷ 2 = 1 if you're not convinced!)

3. Mode (modal value)

- The **mode** (or **modal value**) is simple to identify
- Look for the highest frequency – and then you find the corresponding data value

  In the example above the highest frequency is 15

  Modal number of pets = 1

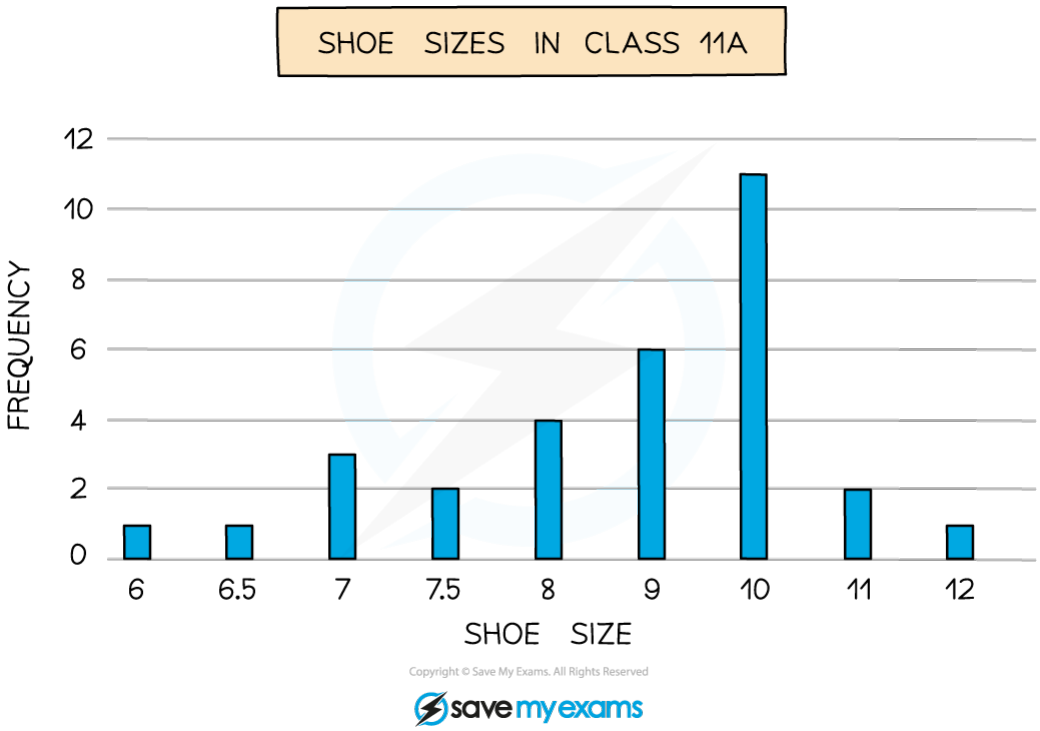
- Don't confuse the number of pets with the frequency

9. Statistics

YOUR NOTES

Worked Example

1. The bar chart shows data about the shoe sizes of pupils in class 11A
   (a) Find the mean shoe size for the class,
   (b) Find the median shoe size,
   (c) Suggest a reason why a shoe shop owner might want to know the modal shoe size of their customers.



9. Statistics

YOUR NOTES ⬇

(a)

1 - Although the data is given in a bar chart this is essentially the same as a table

You should rewrite it as a table and that allows you to add in that extra column for working out

Shoe size	Frequency	Shoe size x Frequency
6	1	6 x 1 = 6
6.5	1	6.5 x 1 = 6.5
7	3	7 x 3 = 21
7.5	2	7.5 x 2 = 15
8	4	8 x 4 = 32
9	6	9 x 6 = 54
10	11	10 x 11 = 121
11	2	11 x 2 = 22
12	1	12 x 1 = 12
Total	31	289.5

Mean = 289.5 ÷ 31 Be careful to get this the right way round!

Mean = 9.3387...

Mean = 9.3 **(to one decimal place)** Round to something sensible

(Note that the mean doesn't have to be an actual shoe size)

(b)

2 - You need to start by finding the position of the median

Position of median = (31 + 1) ÷ 2 = 16th

Median = 9

There are 1 + 1 + 3 + 2 + 4 = 11 values used by the first four rows then the next row with 6 would take you past the 16th value. So the 16th value must be 9.

(c)

A shoe shop manager would want to know the modal size of shoe of his customers as this would be the size of shoes that they are likely to sell most of so he would need to order more of these than the other sizes.

9. Statistics

YOUR NOTES ⬇

9.2.3 CALCULATIONS WITH THE MEAN

Solving problems involving the mean

- Because the mean has a formula it means you could be asked questions that use this formula backwards and in other ways
- Since **Mean = Total of values ÷ Number of values** then it is a formula involving 3 quantities
- Therefore, if you know 2 of these you can find the other one

What may I be asked to do?

- Typical questions ask you to work backwards from a known mean or to combine means for two data sets
- But as this is in the area of problem-solving there may be something unusual that you haven't seen before so you will need to make sure you **understand what the mean is, how it works and what it shows**

1. Working backwards

- For example, the mean of the six data values 5, 7, 2x, 6, 8 and 4x have a mean of 5.4
  Find the value of x
  - This is a matter of setting up an equation in x and solving it

$$\frac{5+7+2x+6+8+4x}{6} = 5.4$$

$$\frac{6x+26}{6} = 5.4$$

$$6x + 26 = 5.4 \times 6$$

$$6x + 26 = 32.4$$

$$6x = 6.4$$

$$x = 1.07 \text{ (to two decimal places)}$$

9. Statistics

YOUR NOTES ⬇

2. Combining two sets of data

- These sorts of questions are generally harder and need more thinking about
- For example, a class of 20 ran a 100m race
  The mean time for the 8 boys in the class was 28.4 seconds
  The mean time for the girls was 32.1 seconds
  Find the mean time for the whole class

$$Class\ mean = \frac{Total\ of\ both\ boys'\ and\ girls'\ times}{Number\ of\ boys\ and\ girls\ in\ total}$$

We know:

There are 20 in the class in total

8 boys in the class and so there are 20 – 8 = 12 girls.

The boys' mean is 28.4 – ie "Boys Total" ÷ 8 = 28.4

The girls' mean is 32.1 – ie "Girls Total" ÷ 12 = 32.1

We do not know:

The boys' total time

The girls' total time

The class' total time

But we can work these out

Boys' total time = 28.4 x 8 = 227.2

Girls' total time = 32.1 x 12 = 385.2

Class' total time = 227.2 + 385.2 = 612.4

And now we can work out the class mean

$$Class\ mean = \frac{612.4}{20} = 30.62 \text{ seconds}$$

Exam Tip

You have used the mean so often in mathematics that you do not normally think of it as a formula. But it is. And as with other work in using formulas you should write down the information you do know, and the information you are trying to find.

9. Statistics

YOUR NOTES ⬇

Worked Example

1. A class of 24 students have a mean height of 1.54 metres.
   Two new students join the class and the mean height of the class increases to 1.56 metres.
   Given that the two new students are of equal height, find their height.

$Mean\ of\ "original"\ class: \frac{Total\ of\ Heights}{24} = 1.56$

$Mean\ of\ "new"\ class: \frac{Total\ of\ Heights\ with\ 2\ new\ students}{24+2} = 1.58$

2 – Two sets of data – before the new students and after

Writing down what we know and what we don't know – using a combination of words and formulas is fine

"Total of heights" = 1.56 x 24 = 37.44

"Total of heights with 2 new students" = 1.58 x 26 = 41.08

Two new heights combined = 41.08 – 37.44 = 3.64

The difference in these totals will be the total of the two new heights

Height of both new students = 3.64 ÷ 2 = 1.82 metres

As both new students have the same height divide by 2 to find their individual heights

9. Statistics

YOUR NOTES ⬇

9.2.4 IQR & RANGE

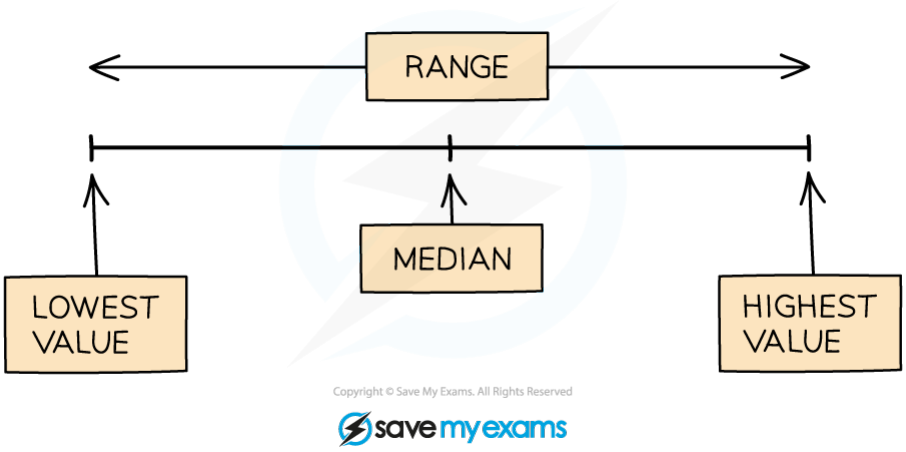
What are IQR and the range?

- The three averages (**mean**, **median** and **mode**) measure what is called **central tendency** – all give an indication of what is typical about the data, what lies roughly in the middle, etc.
- The **range** and **inter-quartile range** (**IQR**) measure how **spread** out the data is
- They only apply to numerical data, and both are easy to work out!

What do I need to know?

1. Range (Hi-Lo)

- This is the difference between the highest value in the data and the lowest value



Range = Hi – Lo

- It is usually meant by "average" – it's like an ideal world where everybody has the same, everything is shared out equally
- It is the **TOTAL** of all the values **DIVIDED** by the **NUMBER OF VALUES**
- For example, find the range of 14, 16, 18, 22

Hi = 22

Lo = 14

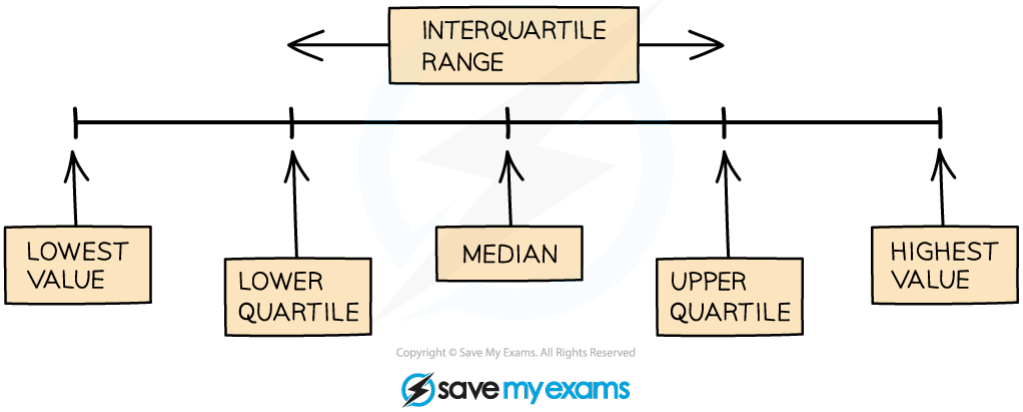
9. Statistics

YOUR NOTES ⬇

Range = 22 – 14 = 8

2. Inter-Quartile Range (IQR)

- This is the difference between the **upper quartile** and the **lower quartile**
- You know the median splits data into two
- Well as their name suggests, quartiles split the data into four



IQR = UQ – LQ

- The **lower quartile (LQ)** is the value **one quarter** of the way along the data
- To find its position we calculate $\frac{n+1}{4}$, where n is the number of data values
- The **upper quartile (UQ)** is the value **three quarters** of the way along the data
- To find its position we calculate $\frac{3(n+1)}{4}$, where n is the number of data values
- The inter-quartile range is then the difference between these

9. Statistics

YOUR NOTES ⬇

For example, find the inter-quartile range of the follow data ...

20, 23, 32, 35, 37, 38, 43, 45, 47, 49, 52, 56, 58, 58, 59

There are 15 values ($n = 15$)

Position of LQ $= \frac{15+1}{4} = 4^{th}$

so LQ = 35

Position of UQ $= \frac{3(15+1)}{4} = 12^{th}$

so UQ = 56

IQR = 56 − 35 = 21

Exam Tip

Remember with the range that you have to do a calculation (even if it is an easy subtraction). It is not good enough to write something like the range is 14 to 22.

Worked Example

9. Statistics

YOUR NOTES ⬇

1. (a) Find the range and the inter-quartile range for the following data

3.4	4.2	2.8	3.6	9.2	3.1	2.9	3.4	3.2
3.5	3.7	3.6	3.2	3.1	2.9	4.1	3.6	3.8
3.4	3.2	4.0	3.7	3.6	2.8	3.9	3.1	3.0

(b) Give a reason why, in this case, the inter-quartile range may be a better measure of how spread out the data is than the range.

(a)

2.8	2.8	2.9	2.9	3.0	3.1	3.1	3.1	3.2	3.2
3.2	3.4	3.4	3.4	3.5	3.6	3.6	3.6	3.6	3.7
3.7	3.8	3.9	4.0	4.1	4.2	9.2			

Values need to be in order, lay them out neatly especially if they do not fit on one line and double check you've not missed any out.

Range = 9.2 – 2.8 — 1 – Range (Hi – Lo)

Range = 6.4

There are 27 values ($n = 27$)

Position of LQ = $\frac{27+1}{4} = 7^{th}$ So LQ = 3.1

Position of UQ = $\frac{3(27+1)}{4} = 21^{st}$ So UQ = 3.7

Notice that if you have worked out the position of the LQ already, the UQ position is just the LQ Position multiplied by 3

IQR = 3.7 – 3.1 — 2 – IQR (UQ – LQ)

IQR = 0.6

(b)

The IQR would be a better measure of spread for these data as the highest value (9.2) is very far away from the rest of the numbers. It could be an outlier.

9. Statistics

YOUR NOTES ⬇

9.3 GROUPED DATA

9.3.1 AVERAGES FROM GROUPED DATA

What is grouped data and why use it?

- Some data for a particular scenario can vary a lot
- For example, the heights of people, particularly if you include a mixture of children and adults
- Because data like height is also **continuous** (essentially data that can be measured) it would be difficult, even using a table, to list every height that gets recorded – also, there is little difference between someone who is 176 cm tall and someone who is 177cm tall
- So we often group data into **classes** but that leads to one important point....

What do I need to know?

- When data is grouped we lose the **raw data**
- With height data this means we might know, how many people have a height of between 150 cm and 160 cm but not the specific heights of those 10 people
- This means we cannot find the actual **mean**, **median** and **mode** from their original definitions – but we can **estimate the mean** and we can also talk about the **class the median lies in** and the **modal class**

1. **Estimating the Mean**

- There is one extra stage to this method compared to finding the mean from tables with discrete data – use the class **midpoints** as our data valueseg. the heights of 25 members of a youth club were recorded and the results are summarised in the table below – estimate the mean height

9. Statistics

YOUR NOTES ⬇

Height, h cm	Frequency
$120 \leq h < 130$	4
$130 \leq h < 140$	5
$140 \leq h < 145$	4
$145 \leq h < 150$	5
$150 \leq h < 160$	3
$160 \leq h < 180$	4

- Note that the **class widths** (group sizes) are not all equal (this is not a problem so do not let it put you off) and be careful with the inequality signs: a height of exactly 130 cm would be recorded in the second row not the first
- As we don't know the original data we use the midpoint of each group – this is the height that is half way between the start and the end of the group
- Usually these are easy to 'see' but you can always work it out if in doubt (eg halfway between 140 and 145 is (140 + 145) ÷ 2 = 285 ÷ 2 = 142.5)
- We then use theses midpoints as the heights for all the people in the 120 ≤ h < 130 class – so we assume that all 4 people in the class will have a height of 125 cm
- This is why the mean will be an estimate – we assume the heights in all the classes 'average out' at the midpoint height

Height, h cm	Frequency	Midpoint
$120 \leq h < 130$	4	125
$130 \leq h < 140$	5	135
$140 \leq h < 145$	4	142.5
$145 \leq h < 150$	5	147.5
$150 \leq h < 160$	3	155
$160 \leq h < 180$	4	165

9. Statistics

YOUR NOTES ⬇

- Now we can proceed as if it were discrete data and multiply, including a total row as well

Height, h cm	Frequency	Midpoint	Frequency x Midpoint
$120 \leq h < 130$	4	125	4 x 125 = 500
$130 \leq h < 140$	5	135	5 x 135 = 675
$140 \leq h < 145$	3	142.5	3 x 142.5 = 427.5
$145 \leq h < 150$	6	147.5	6 x 147.5 = 885
$150 \leq h < 160$	3	155	3 x 155 = 465
$160 \leq h < 180$	4	165	4 x 165 = 660
Total	25	(not needed)	3612.5

- And finally we can find the mean:
  Mean = 3612.5 ÷ 25 = 144.5
  Mean height is 144.5 cm

2. Median

- Rather than find an actual value for the medium you could be asked to find the class in which the median lies
- The process for finding its position is the same as before so for the above example:

  Position of median = (25 + 1) ÷ 2 = 13

  The median is the 13^{th} value

- From looking at the frequency column we can see the 13^{th} value would fall in the 145 ≤ h < 150 class (it is the last value in this class in fact)
  So we would say the **median lies in the 145 ≤ h < 150 class**

3. Modal Class (Mode)

- Similar to finding the median we are only interested in the class the modal value lies within.
- Again using the example above we can see from the table the highest frequency is 6
- So the **modal class is** 145 ≤ h <150

9. Statistics

YOUR NOTES ⬇

Exam Tip

When presented with data in a table it may not be obvious whether you should use the technique below or the one from the previous notes (see Averages from Tables & Charts) but when you see the phrase "**estimate** the mean" you know that you are in the world of grouped (and usually continuous) data so you know to use the method below.

Worked Example

1. The weights of 20 three-week-old Labrador puppies were recorded at a vet's clinic. The results are shown in the table below.

(a) Estimate the mean weight of these puppies

(b) Write down the modal class

Weight, w kg	Frequency
$3 \leq w < 3.5$	3
$3.5 \leq w < 4$	4
$4 \leq w < 4.5$	6
$4.5 \leq w < 5$	5
$5 \leq w < 6$	2

9. Statistics

YOUR NOTES ⬇

(a)

1 - First thing to do is to add columns as necessary and find the midpoints

We can then proceed as above to complete our table

Height, h cm	Frequency	Midpoint	Frequency x Midpoint
$3 \leq w < 3.5$	3	3.25	3 x 3.25 = 9.75
$3.5 \leq w < 4$	4	3.75	4 x 3.75 = 15
$4 \leq w < 4.5$	6	4.25	6 x 4.25 = 25.5
$4.5 \leq w < 5$	5	4.75	5 x 4.75 = 23.75
$5 \leq w < 6$	2	5.5	2 x 5.5 = 11
Total	20		85

Mean = 85 ÷ 20 = 4.25

The mean weight of the puppies is 4.25 kg

(b)

3 – To find the modal class we look for the highest frequency which is 6

The modal class is $4 \leq w \leq 4.5$

9. Statistics

YOUR NOTES ⬇

9.4 CUMULATIVE FREQUENCY

9.4.1 CUMULATIVE FREQUENCY

What does cumulative frequency mean?

- **Cumulative** basically means "adding up as you go along"
- So cumulative frequency means all of the frequencies for the different classes, totalled
- This allows us to draw a cumulative frequency graph, from which we can find useful information such as the **median** and **quartiles**

What do I need to know?

- You will either be asked to **draw** a cumulative frequency graph or **analyse** one
- It is also possible to draw a box plot from a cumulative frequency graph and you could be asked to **compare** the data with another set of data (see Box Plots)

1. Drawing a cumulative frequency graph

- This is best explained with an example
- For example, the times taken to complete a short general knowledge quiz taken by 50 students are shown in the table below:

Time taken, s seconds	Frequency
$25 \leq s < 30$	3
$30 \leq s < 35$	8
$35 \leq s < 40$	17
$40 \leq s < 45$	12
$45 \leq s < 50$	7
$50 \leq s < 55$	3
Total	50

- When asked to find the cumulative frequencies you may be given another table – or the table above will have an extra column on it
- You should be aware of both possibilities although the process is the same

9. Statistics

YOUR NOTES ⬇

Version 1 – an extra column in the original table:

Time taken, s seconds	Frequency	Cumulative Frequency
$25 \leq s < 30$	3	3
$30 \leq s < 35$	8	3 + 8 = 11
$35 \leq s < 40$	17	11 + 17 = 28
$40 \leq s < 45$	12	28 + 12 = 40
$45 \leq s < 50$	7	40 + 7 = 47
$50 \leq s < 55$	3	47 + 3 = 50
Total	50	(not needed)

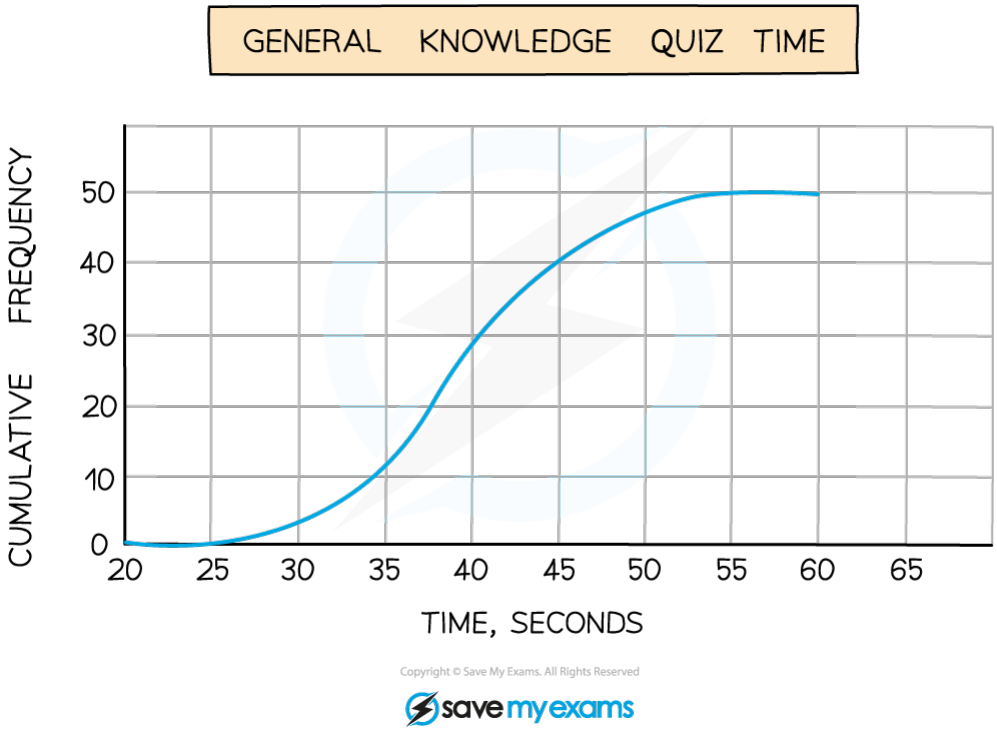
Version 2 – a new table with the class intervals all starting at the same value – the lowest in the data – here this is 25, but will often be 0:

Time taken, s seconds	Cumulative Frequency
$25 \leq s < 30$	3
$25 \leq s < 35$	3 + 8 = 11
$25 \leq s < 40$	3 + 8 + 17 = 28
$25 \leq s < 45$	3 + 8 + 17 + 12 = 40
$25 \leq s < 50$	3 + 8 + 17 + 12 + 7 = 47
$25 \leq s < 55$	3 + 8 + 17 + 12 + 7 + 3 = 50

- Note the cumulative frequencies are found in exactly the same way whichever version you may come across but we've shown two different ways of thinking about it
- Now you have your cumulative frequencies you can draw a graph from it
- The key here is that cumulative frequencies are plotted against the **end** (**upper bound**) of the class interval
- This is because you can't say, for example, you have covered all 11 students (2^{nd} row of table) until you have reached 35 seconds
- After 30 seconds (start) only 3 people had finished the quiz and you cannot tell how many had finished by 32.5 seconds (midpoint)

9. Statistics

YOUR NOTES

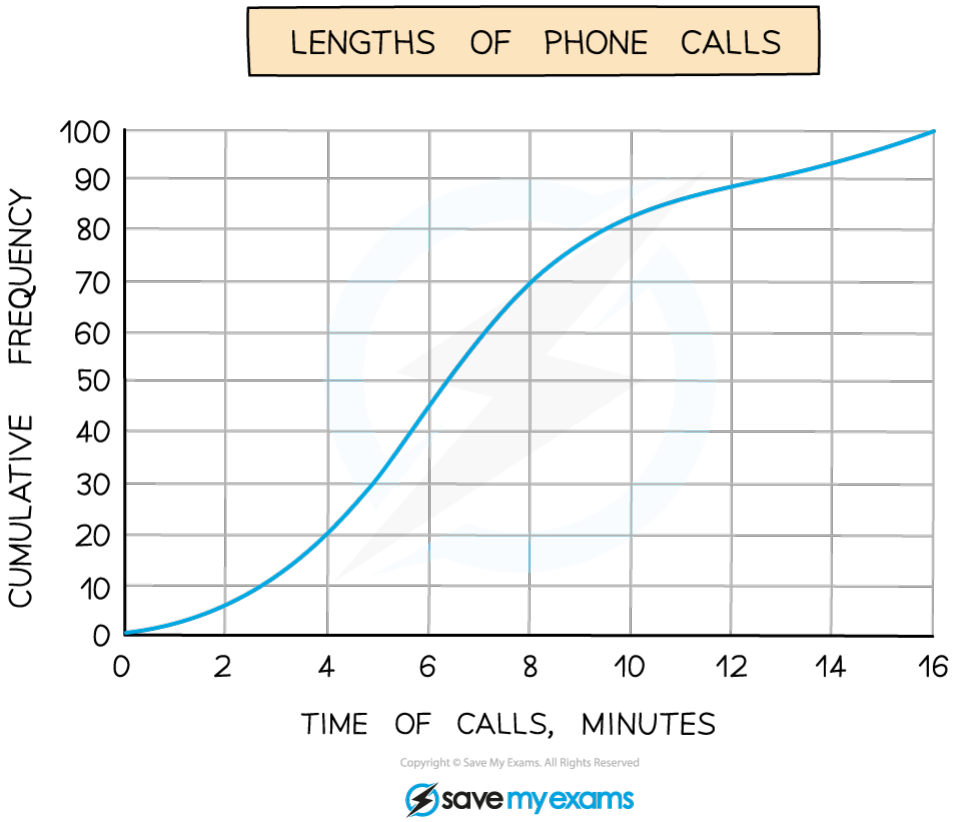


Worked Example

1. A company is investigating the length of telephone calls customers make to its help centre. The company randomly selects 100 phone calls from a particular day and the results are displayed in the cumulative frequency graph below.
   (a) Estimate the median, lower quartile and upper quartile
   (b) The company is thinking of putting an upper limit of 12 minutes on a call – how many of the 100 phone calls would have been beyond this limit?

9. Statistics

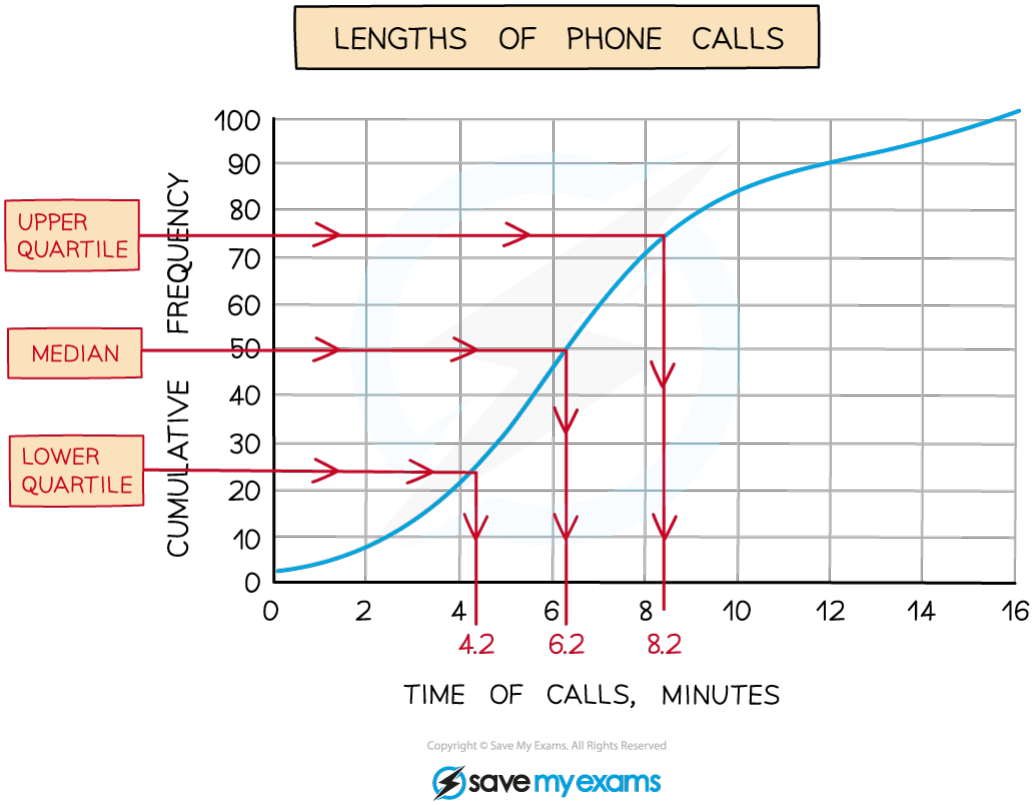
YOUR NOTES



9. Statistics

YOUR NOTES ⬇

(a)



There are 100 pieces of data so to find the median you need the half of this, 50

Starting at 50 on the cumulative frequency axes draw a line across to the graph and down to the time axis and take a reading

Similarly for the lower quartile draw a line from 100 ÷ 4 = 25 and for the upper quartile draw a line across from 100 ÷ 4 x 3 = 75

Median = 6.2 minutes (6m 12s)

Lower Quartile = 4.2 minutes (4m 12s)

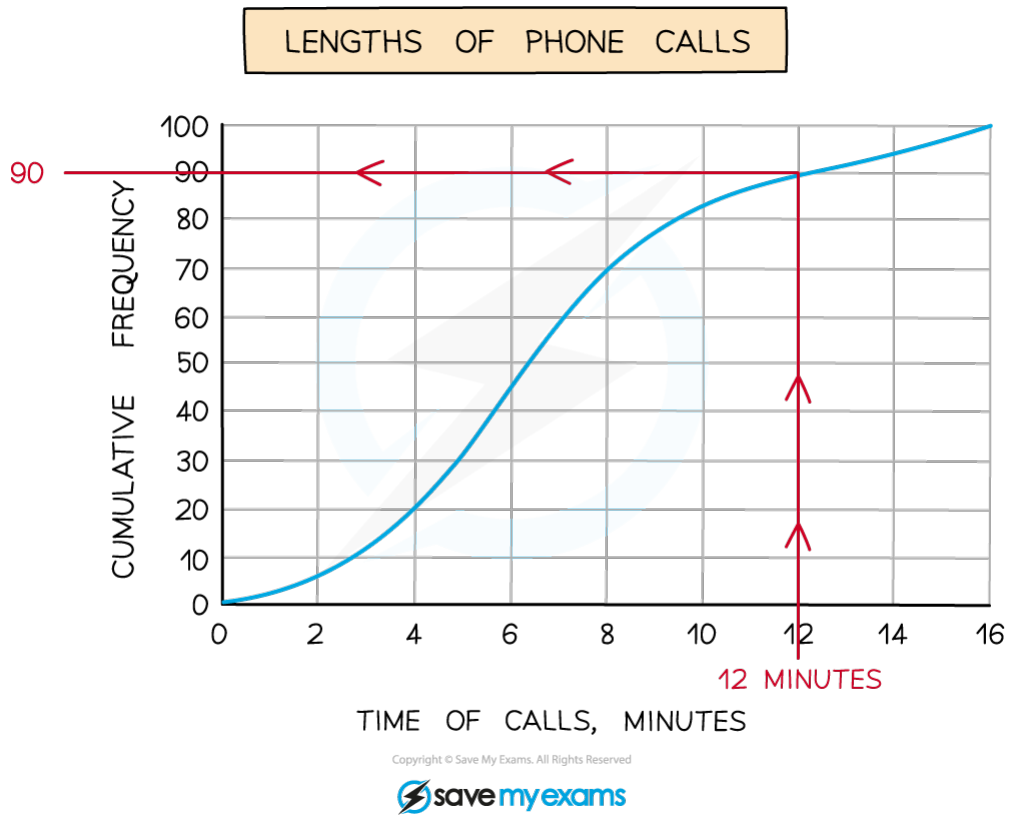
Upper Quartile = 8.2 minutes (8m 12s)

There's no need to convert to minutes and seconds unless asked to by the question

9. Statistics

YOUR NOTES

(b)



Draw a line up from the 12 minute mark on the time axis

Draw a line across and take a reading – in this case 90

100 – 90 = 10

10 calls, out of the 100 the company used, would have been longer than the upper limit of 12 minutes

As the company are cutting off calls **greater** than 12 minutes you need to work out how many calls took longer than this

9. Statistics

YOUR NOTES ⬇

9.4.2 BOX PLOTS

What are box plots and when should they be used?

- **Box Plots** are also known as **Box–and–Whisker Diagrams** (you'll see why below)
- They are used when we are particularly interested in splitting data up into **quartiles**
- Often, data will contain extreme values – consider the cost of a car: there are far more family cars around than there are expensive sports cars
  So if you had 50 data values about the prices of cars and 49 of them were family cars but 1 was a sports car then the sports car's value does not fit in with the rest of the data
- Using quartiles and drawing a box plot allows us to split the data and so we can see what is happening at the low, middle and high points in the data

What do I need to know?

1. Drawing Box Plots

- You need to know five values to draw a box plot:

  Lowest data value

  Lower quartile

  Median

  Upper quartile

  Highest data value

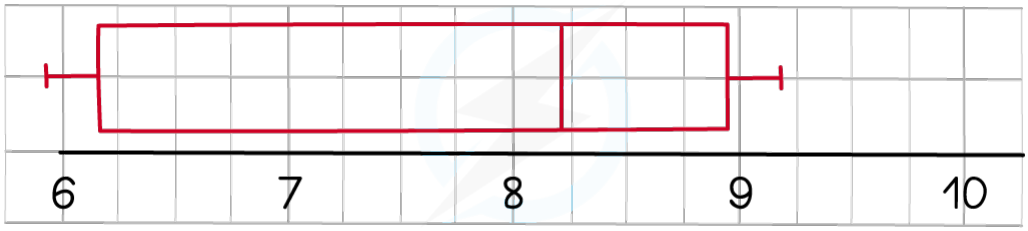
- Conversely, if you are given a box plot you can work out these five values plus other useful statistics like **range** and **inter–quartile range** (IQR)
- Box plots are normally drawn on square or graph paper so you will need to be accurate

9. Statistics

YOUR NOTES ⬇

- For example, given the following information draw a box plot on the graph paper provided

Median	8.2
Lower Quartile	6.2
Upper Quartile	8.9
Lowest Value	5.9
Highest value	9.2



- Plot each point first with a small line – it doesn't matter that they are not listed in order
- The middle three values (lower quartile, median and upper quartile) make the box
- The lowest and highest value make the 'whiskers'

2. Comparing Box Plots

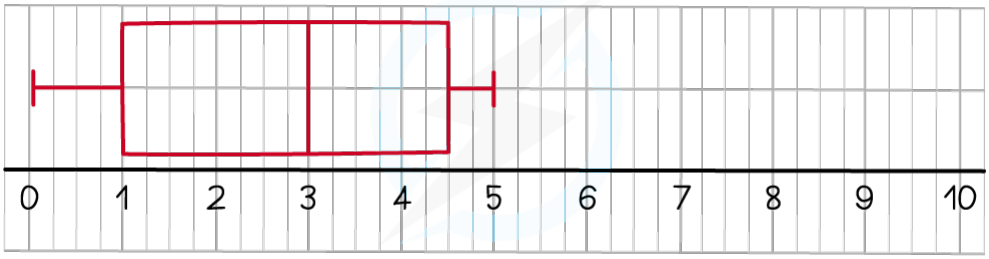
- If you are asked to **compare** box plots (for example between two classes) you should mention at least two things – one about **averages**, ie. **median**, and one about **spread**, ie. **IQR** or **range**

9. Statistics

YOUR NOTES ⬇

Worked Example

1. The box plot below shows the number of goals scored per game by Albion Rovers during a football season.



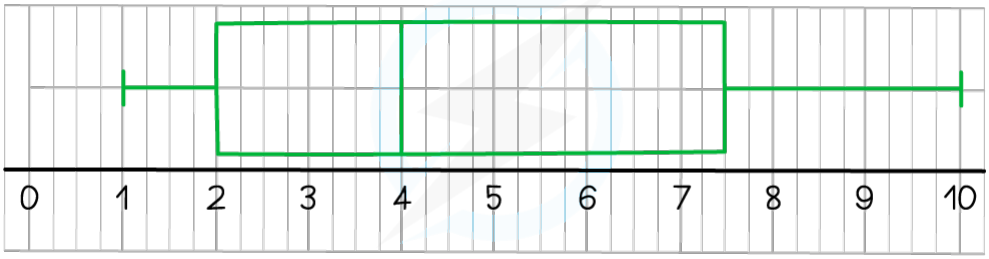
The information below shows the number of goals scored per game by Union Athletic during the same football season.

Median number of goals per game	4
Lower Quartile	2
Upper Quartile	7.5
Lowest number of goals per game	1
Highest number of goals per game	10

(a) Draw a box plot for the Union Athletic data

(b) Compare the number of goals scored per game by the two teams

(a)



9. Statistics

YOUR NOTES ⬇

1 – Draw the box plot by first plotting all five points and then drawing the box around the middle three and whiskers to the outer two

(b)

The median number of goals per game is higher for Union Athletic (4) than Albion Rovers (3). This means that on average, Union Athletic scored more goals per game than Albion Rovers.

2 – This is your first comment about **averages**

Do it in two sentences – one that is just about the maths

The second sentence mentions what the data is showing

The IQR is higher for Union Athletic (5.5) than Albion Rovers (3.5).

This means that Albion Rovers were more consistent with the number of goals they scored per game.

2 – This is your second comment about **spread**.

You have a choice of range or IQR

Both are larger for Union Athletic

Again use two sentences – remember a small range/IQR means less spread out which can be a good thing if you need **consistency** (think golf!)