q) Representing Data.

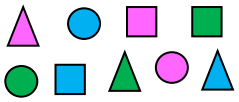
q) Representing Data

Step 1) Sorting into Types

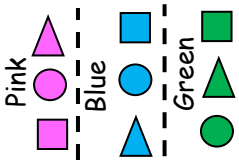
Data is just information about a group of things (or people). The information could be colours, and the things could be cars, so this would be data about the colour of cars. Or it could be the shoe size of people, that is data about people's shoe sizes.

So the information about the cars might be blue, red or green, but it couldn't be 3 metres as that is a distance, not a colour. Shoe size could be 3, 7, 12, but not Germany as that is a country, not a shoe size.

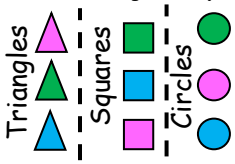
The first thing to learn here is how to sort data into groups of the same type. Here we will take a group of coloured shapes and organise them in terms of two different data types – one is colour, and the other is shape.



Here organised by colour....



& here by shape.



Step 2) Tallying

Tallying is a way of counting how many things we have. You draw a vertical (upright) line for each thing, and every 5th thing you draw a horizontal (sideways) line across the other four.

1 = 𝍸, 2 = 𝍸𝍸, 3 = 𝍸𝍸𝍸, 4 = 𝍸𝍸𝍸𝍸, 5 = 𝍸

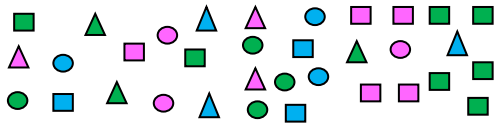
You can see how many there are very quickly and easily.
You can also organise pairs of 5s opposite ways to make 10s.

6 = 𝍸 , 7 = 𝍸, 8 = 𝍸, 9 = 𝍸, 10 = 𝍸

With this system you can count easily, & quickly translate between a tally numbers.

17 = 𝍸 𝍸 and 𝍸 𝍸 𝍸 𝍸 𝍸𝍸𝍸 = 43

In this way we can count groups of different objects (here both by colour and by shape)



By Colour		
Green	Blue	Pink
𝍸 𝍸𝍸𝍸	𝍸	𝍸 𝍸
14	9	11

By Shape		
Circle	Square	Triang.
𝍸	𝍸 𝍸	𝍸
10	15	9

Step 3) Questionaire Design

When designing a questionaire we have to think about whether our questions might effect the answers people give (called biased questions) and how we are going to use the data we collect.

Let's say that you love comedy films and want to see if others do too. If you asked this question...

Q1) I love comedies, what's your favourite type of film?
□ comedy □ horror
□ romance □ action

It would be considered biased as the question itself encourages people to give a particuluar answer. This could be put more neutrally as..

Q1) What is your favourite type of film?
□ comedy □ horror
□ romance □ action

The way the answers are collected is also important. Though open questions are useful, the answers they bring are hard to quantify (put into numbers). So for example...

a) How many magazines do you read each month?

Would lead to many different answers.

But b) How many magazines do you read each month?
□ up to 5 □ some
□10 to 20 □loads

Limits the answers to a manageable number, but how could we be sure where someone who reads 7 magazines might tick?

c) How many magazines do you read each month?
□ 0 □ 1 to 5 □ 6 to 10 □ 11 to 20 □ 11 or more

This is a non-biased question all possible answers are covered, without any overlap.

Step 4) Random Sampling

The POPULATION is all the people or things who's data you are interested in. The might be thousands of them so it is not always possible to collect the data for everyone in the population. So you collect the data from a part of your population and we call this taking a SAMPLE. If you do take data from everyone in your population it is called a CENSUS. With a census your sample is your whole population.

In the UK they take a census collecting lots of different types of data, from of every household in the country, once a decade on years ending with digit 1 (1991, 2001, 2011, and the next will be in 2021). This is the UK national census.

There are different ways to choose which members of your population to choose your data from. If you just picked some by looking through, you might unconsciously choose some that favoured a particular result. It is important that the pieces of data for your sample or chosen randomly.

SYSTEMATIC RANDOM SAMPLING is where you take every 3rd, 4th, 5th etc... piece of data, to give you the correct sample size.

So if we wanted a sample of 4 pieces of data from these 12 pieces of data

3, 6, 8, 14, 18, 34, 37, 89, 123, 156, 345, 539

$\frac{12}{4} = 3$, so we take every 3rd

3, 6, 8, 14, 18, 34, 37, 89, 123, 156, 345, 539

Our sample data are:
8, 34, 123, 539

SIMPLE RANDOM SAMPLING is where you use a random number generator to decide which pieces of data from your ordered list to choose.

On a calculator RAN# gives a random number between 0 &1, so 11Ran# + 1 gives a random number between 1 & 12.

To choose 4 pieces of data, from our twelve with this method, I did I did this 4 times (rounding each to the nearest whole integer) and I got 5, 3, 10, 8 so I take the 3rd, 5th, 8th, & 10th pieces of data.

3, 6, 8, 14, 18, 34, 37, 89, 123, 156, 345, 539
3rd 5th 8th 10th

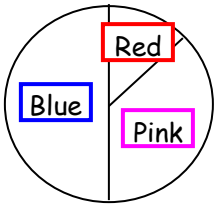
Our sample data are:
8, 18, 89, 156

These are the two main methods for making sure your sample is picked randomly. Later, in step 11 we'll learn more about how to make a sample fair, making sure it represents all the different types of people or things in the population of interest.

Step 6) Ranking from Pie Charts

A pie chart, is a round chart cut into pieces like a pie. The size of each piece of the pie represents the amount of that piece of data within the sample and/or population.

Look at this pie chart representing people's favourite colours.



The largest slice is blue & so blue is the most popular colour.

The smallest slice is red & so red is the least popular colour.

Step 6) Pictograms.

In pictograms we use pictures to represent the number of times a type of data appears. This pictogram represents the number of goals scored by 4 friends in a game. Each star represents 2 goals scored by that person.

Let's work out how many goals Sue scored. She has 5 stars and each one represents 2 goals. 5 x 2 is 10 so Sue scored 10 goals.

Sue = ☆☆☆☆☆ = 5 x 2 = 10 goals

Similarly, for Dom.

Dom = ☆☆☆ = 3 x 2 = 6 goals

The goals for Sal and Jim are a little more compicated as they have some half stars in the diagram. A half star is worth half as many goals as a whole star. Half of 2 goals is 1 goal, so each half star is worth 1 goal.

½ = ½ of 2 = 1 goal

So Jim = ☆☆☆½ = 3½ x 2 = 7 goals

And Sal = ☆½ = 1½ x 2 = 3 goals

Step 7) Bar Charts.

Bar charts are a kind of diagram that represent how many of each type of data you have. The height of each bar tells you how many pieces of that type of data there are (by reading off on a scale on the left).

The FREQUENCY of a type of data, is how many times (how frequently) that data type occurs. So if you had 7 red cars, the frequency of red cars would be 7.

DISCRETE DATA is data which is counted. It can only take particular values. For example colours of cars, or shoe sizes.

CONTINUOUS DATA is data that has to be measured. It can take any value within a particular range.
Let's look at this bar chart of the shoe sizes of a group of people. Because the data is discrete there are gaps between the bars.

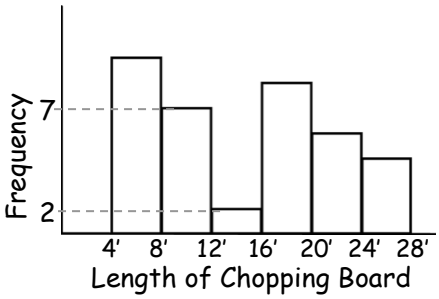
How many had size 4, and how many had size 6 feet?

If we read off the bar for size 4 feet, it gives as a frequency of 5. If we read off the bar for size 6 feet it gives us a frequency of 3.

So 5 people have size 4 feet
And 3 people have size 6 feet.

Lets look at a bar chart of continuous data. The length of some chopping boards in inches' is given in this bar chart. Because we measure lengths, the data is continuous, so there is no gap between the bars.

From the bar chart how many vegetable chopping boards were between 0'&4', 8'&12' and 12'&16'?



Just like for discrete data, we read off the height of the bar we are interested in to find the frequency.
There are 0 boards between 0'&4',
7 boards between 8'&12'
and 2 boards between 12'&16'

Step 8) Stem & Leaf Diagrams

15 people were asked how long they normally spent cooking on a Tuesday. These were their answers in minutes 8, 12, 16, 17, 23, 27, 27, 29, 30, 34, 36, 38, 38, 41 & 47.

This information could be represented in a stem and leaf diagram. The data would be grouped on different parts of a stem. The parts of the stem would be numbers in the 10s 20s, 30s and so on. The leaves would represent the digit.

People's Cooking Times on a Tuesday

5	
4	1 7
3	0 4 6 8 8
2	3 7 7 9
1	2 6 7
0	8

Key.
2 | 7 is 27 mins

You can use stem and leaf diagrams to compare two different sets of data. Let's look at this stem and leaf diagram comparing the times of the same people for cooking on Tuesdays, and Saturdays.

Saturday Times						Tuesday Times				
			8	4	5					
6	6	6	3	2	4	1	7			
	8	6	4	4	3	0	4	6	8	8
		8	6	5	2	3	7	7	9	
				7	1	2	6	7		
					0	8				

Key (Left).
4 | 3 is 34 mins

Key (Right).
2 | 7 is 27 mins

People spent more time cooking on a Saturday, as there are more leaves higher up the stem. A possible reason is that they have more free time for food preparation as many people don't work at weekends.

Step 9) Averages & Range

The RANGE of a group of data, tells you how spread out the data is. It is found by subtracting the smallest piece of data from the largest. In other words it is the distance between the smallest and biggest piece of data.

In Jobo School, the tallest person in year 8 is 174cm, and the shortest person is 132cm. So the range of heights in year 8 is 174 – 132 = 42cm.

In the whole school the tallest person is 189cm and the shortest is 126cm. So the range of heights in the whole of Jobo School is 189 – 126 = 63cm.

As you might expect the range of heights from one year group is less than the range of heights of the whole school. School age people of the same age are likely to be more similra in height than schooe people from different year groups.

RANGE = highest – lowest
It is a measure of how spread the data is.

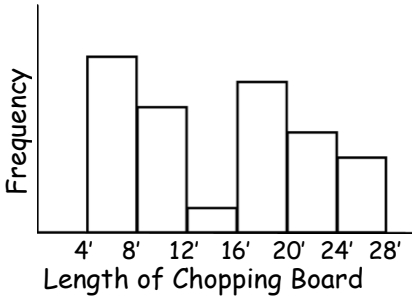
AVERAGE is a measure of how big the data is. It is trying to find a typical value for a group of data.

There are 3 different types of avergae.

The mode (or modal average) is found by taking the most commonly occurring data value.

So for data set 3, 4, 5, 5, 5, 6, 7, 7, 20
5 occurs the most (3) times, so the mode is 5.

Using the example of our bar charts from the above set.



Here the modal group is 4' to 8' chopping boards (as it is the tallest bar and so has the highest frequency).

The other two types of average are called mean and median. They are the most commonly used averages and have different advantages and disadvantages. Nethier is the "best" but one might be better for a particular situation.

The mean average is found by adding up all the data, and dividing by how many pieces of data there are.

Let's work with this set of data: 4, 2, 7, 5, 3, 7 & 14

To recap the range is 14 – 2 = 12 & the mode is 7 (it occurs twice).

To find the mean firstly add up the numbers

$$4 + 2 + 7 + 5 + 3 + 7 + 14 = 42$$

There are 7 numbers, so you then divide by 7.

$$\frac{42}{7} = 6$$

So the mean is 6.

The median is the middle number (when the data is in order).

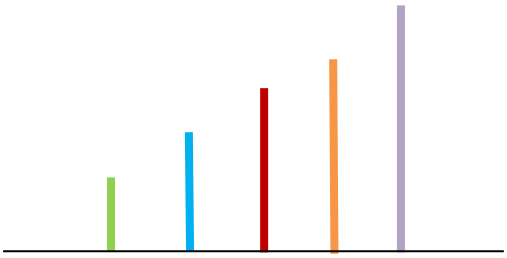
So firstly let's put the numbers in order:
2, 3, 4, 5, 7, 7, & 14

We then cross of numbers each end until we find the middle no.

~~2~~, ~~3~~, ~~4~~, ⑤, ~~7~~, ~~7~~, & ~~14~~

So the median here is 5.

Lets explore the mean and median a little more. Let's look at the length of some lines.



The median average is the length of the red line.

If we made the purple (longest) line even longer, this would not effect the median. Even if me made the purple line 10 miles high. The median would still just be the length of the red line.

Similarly if we made the green line so short you could barely see it, then the red line would still be the median.

But the mean would be effected by making the purple line longer (or the green one shorter).

Because if we made the purple line longer, the total length of all the lines (the added data) would be more so the mean when dividing it by 5 (as there's 5 lines) would be more.

So the median average has the advantage of being quite stable, and is not warped by very large or very small values at either end of the data. The mean has the advantage of taking into account the exact value of every piece of data.

It is also interesting to note, that if we increased the length of the longest (purple) line the range would also get larger. The range is a very quick an easy measure of spread, but has the disadvantage of being very warped by just one exceptionally large (or small) piece of data. There is another measure of spread called inter-quartile range which isn't effected in this way. We will learn about inter-quartile range in step 15.

Step 10) Pie Charts

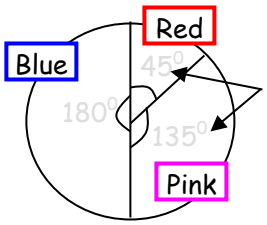
To create pie charts we need to use the fact that there are 360^0 in a full turn, and find the correct fraction of that angle for each data type.

In a survey, 200 people were asked what their favourite colour was. 25 preferred red, 75 pink, & 100 blue.

To create pie chart of this we need to find what fraction of the pie each colour represents. We then need to find that fraction of the 360^0 and measure that on the pie chart.

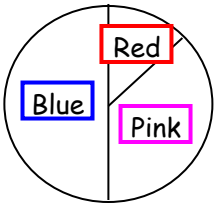
Colour	Freq.	Calculation	Angle
Red	25	$\frac{25}{200}\times 360$	45^0
Pink	75	$\frac{75}{200}\times 360$	135^0
Blue	100	$\frac{100}{200}\times 360$	180^0
Total	200		

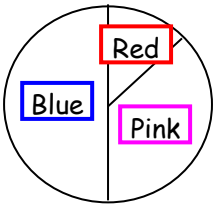
For example, 100 people preferred blue. That means that $\frac{100}{200}$ of the people said blue. This is $\frac{1}{2}$ the people. So we find $\frac{100}{200} of 360°$ which is 180^0.



But you don't have to label the angles.

So the actual pie chart will look like this:

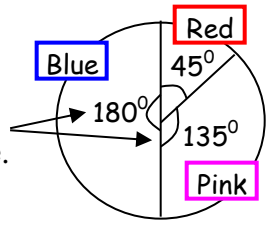




We'll now use the same example in reverse order to show how we could find the frequencies from a pie chart.

200 people were asked their favourite colour. This pie chart shows the results. Find how many preferred each colour.

Measure the angles & put them in a table.



Colour	Angle	Calculation	Freq.
Red	45^0	$\frac{45}{360}\times 200$	25
Pink	135^0	$\frac{135}{360}\times 200$	75
Blue	180^0	$\frac{180}{360}\times 200$	100

Lets look at he red part. The angle fo the red part is 45^0. This represents 45^0 out of 360^0. So the red part is $\frac{45}{360}$ of all the data. As there are 200 pieces of data we find $\frac{45}{360}$ of 200, which is $\frac{45}{360}\times 200 = 25$.
So 25 people preffered red.

Step 11) Stratified Sampling

A strata is like a layer of your polution. For example, if you were looking at a school, year 7 would be one layer, or strata, year 8 another, and so on through to year 13. You could also divide the school into three strata by gender, girls, boys, and other genders.

Stratified sampling is where you ensure the relative size of your population is the same relative size of your sample. So if ¾ of your population had Ethiopian heritage, you would make sure your sample group was of ¾ Ethiopian heritage.

Lets say we had a school with 750 students, with 113 in year7, 204 in year 8, 157 in year 9, 139 in year 10, and 137 in year 11. and we want to choose a sample of 50 students stratified by year group.

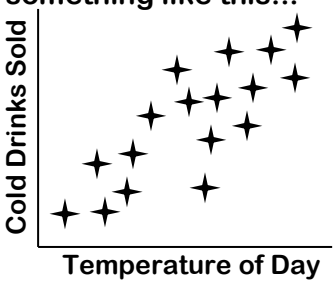
Year	Size	Sample Proportion	Calculation	Number in Sample of 50
7	113	$\frac{113}{750}$	$\frac{113}{750} \times 50$	8
8	204	$\frac{204}{750}$	$\frac{204}{750} \times 50$	14
9	157	$\frac{157}{750}$	$\frac{157}{750} \times 50$	10
10	139	$\frac{139}{750}$	$\frac{139}{750} \times 50$	9
11	137	$\frac{137}{750}$	$\frac{137}{750} \times 50$	9

For example, there are 157 year 9s which means that $\frac{157}{750}$ths of our sample should be in year 9. So we need to find $\frac{157}{750} of 50$ which is $\frac{157}{750} \times 50 = 10$ people.

Step 12) Scatter Diagrams & Correlation.

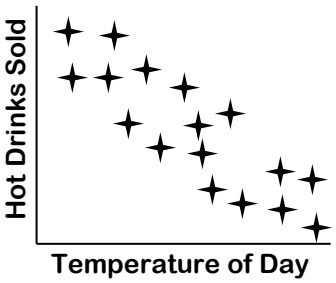
A scatter diagram is a way of comparing to pieces of data from the same sample, to see if there is a relationship. If there is a relationship you say that the two pieces of data are correlated (there is a correlation).

For example on a hot day you would expect to sell more hot drinks and less cold drinks. If you took a sample of days, and plotted the temperature against the number of cold drinks sold it might look something like this...



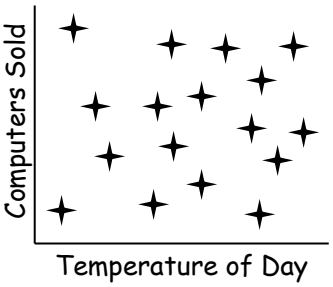
Although the crosses don't forma perfect straight line, never the less you can see a clear relationsip, that the hotter the day the more cold drinks were sold. We call this positive correlation, where one data type increases, the other increases too.

Negative correlation is when as one data type increases, the other decreases. Let's look at a scatter diagram comparing the temperatuure of the day and the number of hot drinks sold.



You can see that there is a negative relationship – the hotter the day, the less hot drinks are sold. So it looks like for this population, there is a negative correlation between temperature and the number of hot drinks sold.

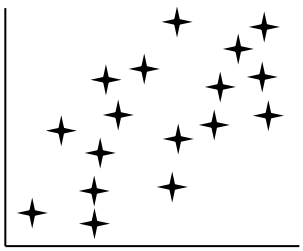
Sometimes we find that there is no cerrelation at all. For example you wouldn't expect the temperature of a day to effect the number of computers sold (in the way it effects hot or cold drink sales). Let's take a look...



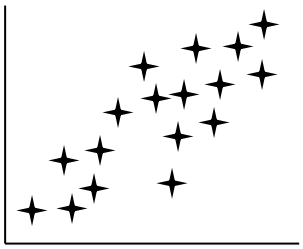
Step 13) Predictions from Scatter Diagrams.

With scatter diagrams you can have a weak correlation, a strong correlation or even a perfect correllation.

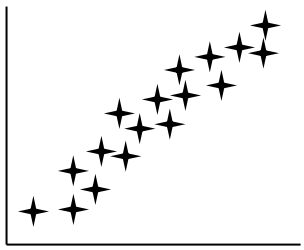
Weak Correlation



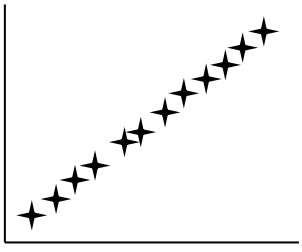
Correlation



Strong Correlation



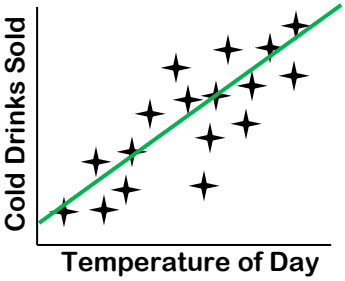
Perfect Correlation



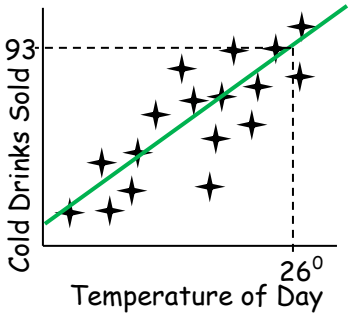
With a perfect correlation, the crosses are already in a perfect straight line and you could easily draw this line in. For a strong correlation you could easily imagine where the line would go if it was a perfect

correlation. This is called a line of best fit. The weaker the correlation the more imagination is needed and the less accurate the line of best fit is, but never-the-less you can draw one!

For example we could draw a line of best fit on the day temperature correlating (positively) with the number of cold drinks sold from step 12.



We could now use this line to make predictions. Let's say we wanted to predict how many cold drinks would be sold on a day that was 26°C. We can read off the temperature axis at 26°C until we reach our line of best fit and then read along to the number of cold drinks axis.



Using this data we would predict that 93 drinks would be sold on a day whose temperature was 26°C.

Step 14) Estimating Mean & Median from Grouped Data.

The table below shows the number of pieces of fruit eaten each week by a sample of 14 people.

Fruit	Frequency
1-5	6
6-10	7
11-20	4

We can't calculate the mean exactly as we don't know exactly how many pieces of fruit each person ate. However, we do know, for example, that 6 people ate somewhere between 1 and 5 pieces of fruit. The best we can do is guess that all 6 of these people ate 3 pieces of fruit this is the midpoint between 1 and 5 pieces. This would mean that 18 pieces of fruit in total were eaten between the 6 people in the 1-5 pieces group. We do this for all the other groups.

Fruit	Group Midpoint	Frequency	Midpoint × Frequency
1-5	3	6	18
6-10	8	7	56
11-20	15.5	4	62
Total		14	136

Mean $\approx \frac{136}{14} = 9.7$ pieces of fruit (1dp)

If our data is continuous, we estimate it in exactly the same way, by pretending that every person's data was in the exact middle (midpoint) of their group.

This table shows the width of a sample of 100 stickers (in mm) from a book with 1000s of stickers. Here's how we would estimate the mean width of a sticker from this book.

Sticker Width (mm)	Group Midpoint	Frequency	Midpoint x Freq.
$0 \leq x < 10$	5	5	25
$10 \leq x < 20$	15	12	180
$20 \leq x < 35$	27.5	17	467.5
$35 \leq x < 50$	42.5	31	2015
$50 \leq x < 65$	57.5	21	1207.5
$65 \leq x < 100$	82.5	14	1155
Total		100	5050

Mean $\approx \frac{5050}{100} = 50.5$ mm

We could also estimate the median from this group. The median is the middle piece of data. So it would be the 50th piece of data from the 100 stickers. There are 5 in the $0 \leq x < 10$ group, then 12 in the $10 \leq x < 20$ group (that's 17 so far, leaving 33 more to count), then 17 in the $20 \leq x < 35$ group (that's 34 so far – leaving 16 more to go), then 31 in the $35 \leq x < 50$ group. So the 50th piece of data, would be the 16th piece in the $35 \leq x < 50$ group. We can find a good estimate by pretending the 31 bits of data are in order in the group, and that the 16th bit is $\frac{16}{31}$ of the way through the group. The width of the group is 50 – 35 which is 15. So we add $\frac{16}{31} \times 15$ onto 35 (the bottom of the group). This gives an estimate for the median as

$35 + \frac{16}{31} \times 15 = 42.7$mm (1dp)

Step 15) Cumulative Frequency & Interquartile Range.

At the end of step 14 we looked at how the median is half way through all our pieces of data (when they are in order). We can also imagine the piece of data ¼ of the way thorugh and call it the lower quartile. In a similar way we can describe the piece of data ¾ of the way through as the upper quartile.

We do this by plotting upper boundary of each group with the sum of all the frequencies of data up to that point. This is called the cumulative frequency, as it is

the sum (the accumulation) of all the frequencies up to that point.

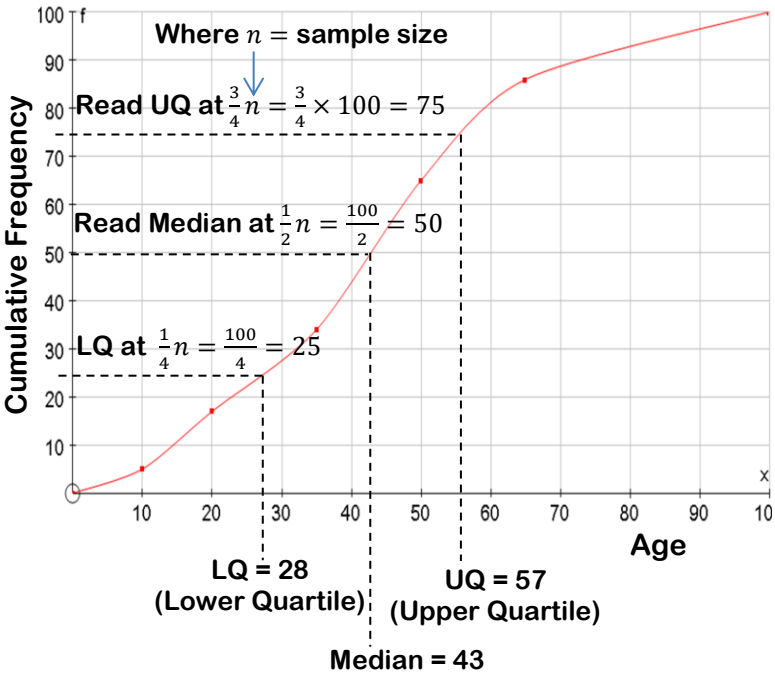
Look at the following data for the ages of 100 people.

Age of Person	Frequency	Cumulative Frequency
$0 \leq x < 10$	5	5
$10 \leq x < 20$	12	17
$20 \leq x < 35$	17	34
$35 \leq x < 50$	31	65
$50 \leq x < 65$	21	86
$65 \leq x < 100$	14	100

Age of Person	Freq-uency		Cumulative Group	Cumulative Frequency
$0 \leq x < 10$	5	→	$0 \leq x < 10$	5
$10 \leq x < 20$	12	→	$0 \leq x < 20$	17
$20 \leq x < 35$	17	→	$0 \leq x < 35$	34
$35 \leq x < 50$	31	→	$0 \leq x < 50$	65
$50 \leq x < 65$	21	→	$0 \leq x < 65$	86
$65 \leq x < 100$	14	→	$0 \leq x < 100$	100

TOP

We then plot these cumulative frequencies against the cumaltive group upper bound – and this gives us a graph we can read off the lower quartile which is ¼ throught the data (here 25th piece of data), the median which is ½ way through the data (here 50th piece of data) and the upper quartile which is ¾ of the way through the data (here 75th piece of data).



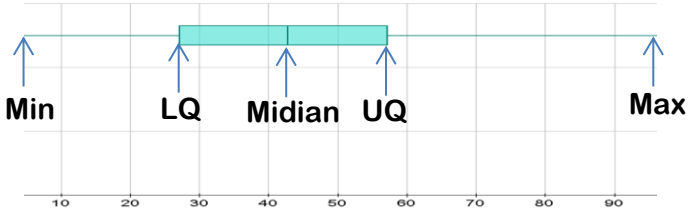
We saw in step 9 that the range is not very stable as if the last or first piece of data change, then so does the range. The interquartile range is much more stable, this is the difference betweent the upper quartile and the lower quartile.

In the above example our interquartile range (IQR)

IQR = UQ – LQ = 57 – 28 = 29

Lots of pieces of data could change, but the IQR range would only be changed when enough pieces of data change so as to put a different piece of data ¼ and/or ¾ of the way through the list of ordered data.

The quartiles and the median can be sumerised on a box and whisker plot as below....



You always need a scale below it, and you can use this to compare two sets of data. For example with a box and whisker plet you could quickly compare the data for heights of year 7s and year 8s, comparing the quartiles and the medians, as well as the spread (both range & IQR) of the data.

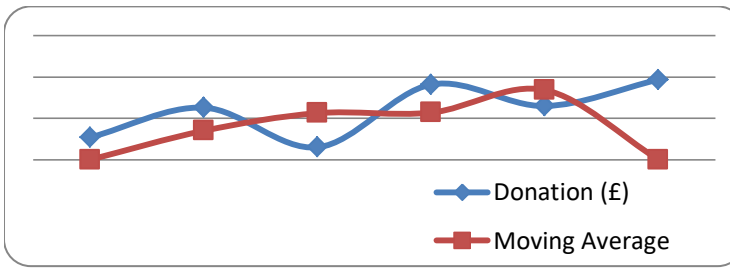
Step 16) Moving Averages.

Calculate the three point moving average for the charitable donations given between 2002 & 2007 in the table on the right, rounded to the nearest pound.

Year	2002	2003	2004	2005	2006	2007
Dona-tion	£27	£63	£15	£91	£65	£97
3 Point Moving Average	/	$\frac{27+63+15}{3}$	$\frac{63+15+91}{3}$	$\frac{15+91+65}{3}$	$\frac{91+65+97}{3}$	/
	/	£35.00	£56.33	£57.00	£84.33	/
	/	£35	£56	£57	£84	/

We can plot these on a graph called a time series.

A time series for the charitable donations between 2002 & 2007.



The relationship described by a time series is called a trend. The trend in this example, is a gradual increase in the charitable donations between 2002 & 2007.

Step 17) Histograms.

A histogram is a special type of bar chart where it isn't the height of each bar that represents the frequency, but the area of each bar.

This is useful, because with grouped data, a wider group width will make it likely to have a higher frequency – but a bar chart will just show this wider group as if it has a higher frequency but per unit width which is not fair!

100 people at a concert are picked at random, there ages are counted in the age classes below.

Age of Person	Frequency
$0 \leq x < 10$	**5**
$10 \leq x < 20$	**12**
$20 \leq x < 35$	**30**
$35 \leq x < 50$	**45**
$50 \leq x < 70$	**34**
$70 \leq x < 100$	**21**
Total	**100**

If we plot these in a bar chart it looks like there are more people between 35 and 50 than any other age. But the group is wider than some and narrower than others, and this affects the frequency you would expect.

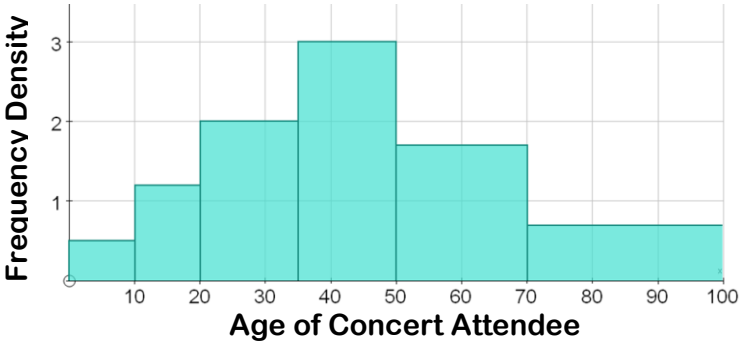
With a histogram, we allow for this by dividing the frequency of each group by the group (or class) width. We call this the frequency density.

$$Frequency\ Density = \frac{Frequency}{Class\ Width}$$

Age of Person	Frequency	Class Width	Frequency Density
$0 \leq x < 10$	**5**	**10**	$\frac{5}{10} = 0.5$
$10 \leq x < 20$	**12**	**10**	$\frac{12}{10} = 1.2$
$20 \leq x < 35$	**30**	**15**	$\frac{30}{15} = 2$
$35 \leq x < 50$	**45**	**15**	$\frac{45}{15} = 3$
$50 \leq x < 70$	**34**	**20**	$\frac{34}{20} = 1.7$
$70 \leq x < 100$	**21**	**30**	$\frac{21}{30} = 0.7$
Total	**100**		

We then plot the data just like a bar chart, but with the vertical axis being frequency density instead of frequency.

A histogram for the ages of 100 people at the concert.



When you have a histogram, the frequency is determined by the AREA of each bar.

$$Frequency = Frequency\ Density \times Class\ Width$$

Looking at the 3rd bar of this histogram. The class goes from age 20 to age 35, so the class width is 15. The frequency density is 2.

$$Frequency = 2 \times 15 = 30$$

So the 3rd bar represents 30 people (which we can see from the original table is correct).