

YEAR 8 - REASONING WITH DATA...

Measures of location

@whisto_maths

What do I need to be able to do?

By the end of this unit you should be able to:

- Understand and use mean, median and mode
- Choose the most appropriate average
- Identify outliers
- Compare distributions using averages and range

Keywords

Spread: the distance/ how spread out/ variation of data

Average: a measure of central tendency – or the typical value of all the data together

Total: all the data added together

Frequency: the number of times the data values occur

Represent: something that shows the value of another

Outlier: a value that stands apart from the data set

Consistent: a set of data that is similar and doesn't change very much

Mean, Median, Mode

The Mean

A measure of average to find the central tendency... a typical value that represents the data

24, 8, 4, 11, 8

Find the sum of the data (add the values) 55

Divide the overall total by how many pieces of data you have $55 \div 5$

Mean = 11

The Median

The value in the center (in the middle) of the data

24, 8, 4, 11, 8

Put the data in order

4, 8, 8, 11, 24

Find the value in the middle

4, 8, 8, 11, 24

Median = 8

NOTE: If there is no single middle value find the mean of the two numbers left

The Mode (The modal value)

This is the number OR the item that occurs the most (it does not have to be numerical)

24, 8, 4, 11, 8

This can still be easier if the data is ordered first

4, 8, 8, 11, 24

Mode = 8

Choosing the appropriate average

The average should be a representative of the data set – so it should be compared to the set as a whole - to check if it is an appropriate average

Here are the weekly wages of a small firm

£240	£240	£240	£240	£240
£260	£260	£300	£350	£700

Which average best represents the weekly wage?

The Mean = £307

The Median = £250

The Mode = £240

Put the data back into context

Mean/Median – too high (most of this company earn £240)

Mode is the best average that represents this wage

It is likely that the salaries above £240 are more senior staff members – their salary doesn't represent the average weekly wage of the majority of employers

Identify outliers

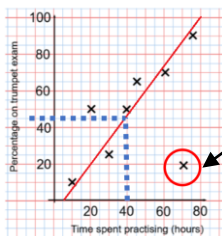
Outliers are values that stand well apart from the rest of the data

Outliers can have a big impact on range and mean. They have less impact on the median and the mode

Sometimes it is best to not use an outlier in calculations

Height in cm
152 150 142 158 182 151 153 149 156 160 151 144

Where an outlier is identified try to give it some context. This is likely to be a taller member of the group. Could it be an older student or a teacher?



Outliers can also be identified graphically e.g. on scatter graphs

Comparing distributions

Comparisons should include a statement of average and central tendency, as well as a statement about spread and consistency

Here are the number of runs scored last month by Lucy and James in cricket matches

Lucy: 45, 32, 37, 41, 48, 35

James: 60, 90, 41, 23, 14, 23

Lucy

Mean: 39.6 (1dp), Median: 38, Mode: no mode, Range: 16

James

Mean: 41.8 (1dp), Median: 32, Mode: 23, Range: 76

James has two extreme values that have a big impact on the range

"James is less consistent than Lucy because his scores have a greater range. Lucy performed better on average because her scores have a similar mean and a higher median"

Interpreting & Representing Data

What do I need to be able to do?

By the end of this unit you should be able to:

- Set up a statistical enquiry
- Draw and interpret pictograms, bar charts and vertical line charts
- Draw and interpret multiple bar charts
- Draw and interpret pie charts
- Represent and interpret grouped quantitative data
- Compare distributions

Keywords & definitions:

Hypothesis: an idea or question you want to test

Sampling: the group of things you want to use to check your hypothesis

Primary Data: data you collect yourself

Secondary Data: data you source from elsewhere e.g. the internet/newspapers/ local statistics

Discrete Data: numerical data that can only take set values

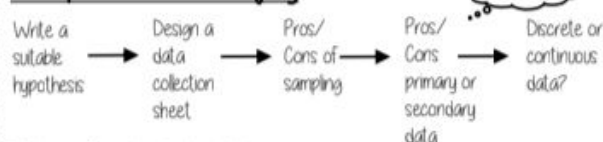
Continuous Data: numerical data that has an infinite number of values (often seen with height, distance, time)

Spread: the distance/ how spread out/ variation of data

Average: a measure of central tendency – or the typical value of all the data together

Proportion: numerical relationship that compares two things

Set up a statistical enquiry



Features of a data collection sheet

Data Title	Tally	Frequency
Grouped or ungrouped categories		Total number of that group observed

Choosing the right chart:

- Comparing values - bar charts, line graphs, pie charts
- Observing trends over time - Line graphs
- Looking at distributions - Pie charts, bar charts, pictograms
- Showing proportions - Pie charts and/or composite bar charts

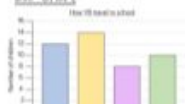
Pictograms, bar and line charts

Pictogram

Language	Frequency
French	4
Spanish	3
German	2

- Need to remember a key
- Visually able to identify mode

Bar Chart



- Gaps between the bars
- Clearly labelled axes
- Scale for the axes
- Title for the bar chart
- Discrete Data

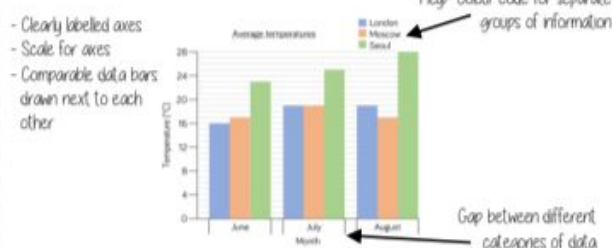
Line Chart



- Gaps between the lines
- Clearly labelled axes
- Scale for the axes
- Discrete Data

Multiple Bar chart

Compares multiple groups of data



Draw and interpret Pie Charts

Type of pet	Dog	Cat	Hamster
Frequency	32	25	3

$\frac{32}{60}$ "32 out of 60 people had a dog"

This fraction of the 360 degrees represents dogs

$$\frac{32}{60} \times 360 = 192^\circ$$



Use a protractor to draw
This is 192°

Remember a circle has 360°

There were 60 people asked in this survey (Total frequency)

Multiple method
As 60 goes into 360 – 6 times
Each frequency can be multiplied by 6 to find the degrees (proportion of 360)

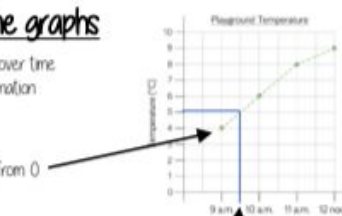
Represents quantitative, discrete data

Draw and interpret line graphs

- Commonly used to show changing over time
- The points are the recorded information and the lines join the points

Line graphs do not need to start from 0

More than one piece of data can be plotted on the same graph to compare data

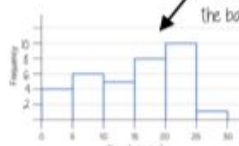


It is possible to make estimates from the line
e.g. temperature at 9.30am is 5°C

Grouped quantitative data

Time (minutes)	Frequency
$0 \leq t < 5$	4
$5 \leq t < 10$	6
$10 \leq t < 15$	5
$15 \leq t < 20$	8
$20 \leq t < 25$	10
$25 \leq t < 30$	1

"More than or equal to 25 and less than 30 minutes"



The use of inequalities shows that this will be a frequency diagram

This is a frequency diagram
There are no gaps between the bars

Grouping the data is useful if there is a large spread of data to begin with

Find and interpret the range

The range is a measure of **spread**

A smaller range means there is less variation in the results – it is more consistent data

A range of 0 means all the data is the same value

Shop 1 has the smallest range – this indicates it has a more consistent flow of customers each week

Difference between the biggest and smallest values



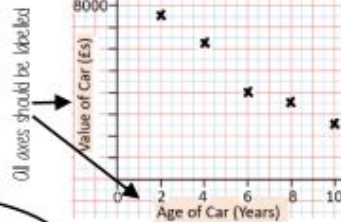
Draw and interpret a scatter graph

Age of Car (Years)	2	4	6	8	10
Value of Car (£)	7500	6250	4000	3500	2500

- This data may not be given in size order
- The data forms information pairs for the scatter graph
- Not all data has a relationship

"This scatter graph shows as the age of a car increases the value decreases"

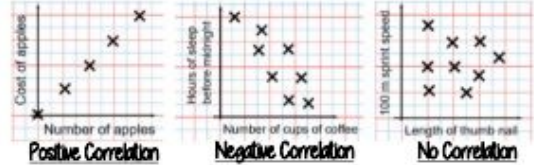
The link between the data can be explained verbally



All axes should be labelled

The axis should fit all the values on and be equally spread out

Linear Correlation



As one variable increases so does the other variable

As one variable increases the other variable decreases

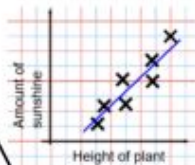
There is no relationship between the two variables

The line of best fit

The Line of best fit is used to make estimates about the information in your scatter graph

Things to know:

- The line of best fit **DOES NOT** need to go through the origin (The point the axes cross)
- There should be approximately the same number of points above and below the line (It may not go through any points)
- The line extends across the whole graph



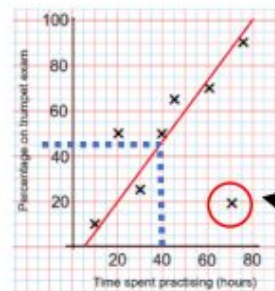
It is only an estimate because the line is designed to be an average representation of the data

It is always a **straight line**

Using a line of best fit

Interpolation is using the line of best fit to estimate values inside our data point

e.g 40 hours revising predicts a percentage of 45



Extrapolation is where we use our line of best fit to predict information outside of our data

This is not always useful – in this example you cannot score more than 100%. So revising for longer can not be estimated

This point is an **"outlier"** It is an outlier because it doesn't fit this model and stands apart from the data

Retrieval Quiz

Answers