

Analysing data

Types of data

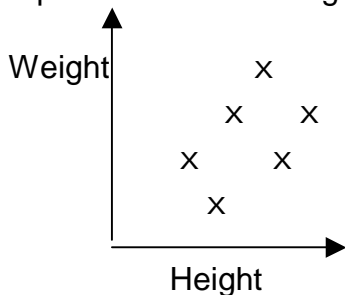
- σ Qualitative - data that is sorted according to type or category.
e.g. Flavour of crisps/type of job/favourite washing powder/eye colour
This data lends itself to representation by a pie chart or bar chart with the horizontal axis labelled with all the different types.
- δ Quantitative - data that is sorted according to a numerical value
e.g. Number of people on a bus/height of person

The number of people on a bus is a **discrete variable** taking *countable* values

The height or weight of a person is a **continuous variable** taking *measurable* values

Analysing the relationship between two variables

Each person can have his or her **height** and **weight** recorded. A cross on a scatter diagram can represent both the height and the weight. Several crosses represent several people.



A **hypothesis** is an assumption made about the data

We could assume that "as the **height** increases so does the **weight**".

To prove our hypothesis we could look at the whole **population** i.e. all pupils

The shortage of time and money means that we usually take a **sample**

Choosing the sample size

With 1500 pupils in a school, a 3% **simple random sample** would involve 45 pupils being chosen

But who will be the chosen few?

Give every pupil a number, put the numbers on paper, roll up the papers, put them in a hat and draw out 45. Nothing wrong with that! This is known as **simple random sampling**.

There is a clever function in EXCEL written **rand()*1500** this should give us a number between 0 and 1500. Round up to get our sample student! Pressing the F9 key in Excel will automatically change the numbers to a new set of random numbers.

574.1427	88.1159	203.268	111.9572	349.0232
575	89	204	112	350

The **Rand** button on your calculator will generate a random number between 000 and 999.

Multiplying this by 1500 will give a random number between 0 and 1500. Again, round up to get your sample student.

A table of random numbers can easily be found, and used in a similar way to get your sample.

40 85 03 89 17 14 32 13 17 51

12 34 79 10 50 40 63 ...the numbers are arranged on a page like this to be easy on the eye.

They may be used individually: 4 0 8 5 0 3 each with a probability of $\frac{1}{10}$

In pairs: 40 85 03 each with a probability of $\frac{1}{100}$

In threes: 408 503 891... giving a three digit number between 000 and 999 with probability $\frac{1}{1000}$

If we need a number between 1 and 1500. Use four digit random numbers, which will be between 0000 and 9999

The first number from the table will be **4085** too big? **Do we just discard it?**

No! Too many numbers will be wasted. The numbers first need to be sectioned into groups of 1500.

0001 - 1500 1501 - 3000 3001 - 4500 9000 Now only numbers above 9000 will be discarded.

The number **4085** falls into the third group, which contains the 1500 numbers 3001 to 4500
4085 is the 1085th number (4085 - 3000) in this group. Choose student number 1085 for your sample.

Alternatively, take the random number (**4085**), divide by 1500 and take the remainder (1085). Ignore the number 0000 if it appears and also ignore numbers above 9000.

Simple random sampling will give you a sample from the whole population: the pupils in the school.

Quota sampling allows you to stop when you have reached your **quota**. This will be the number you require your sample to be but if you take the first 45 pupils, the others won't have a chance of being included. Anyway, you would only get a sample of year sevens unless you scramble the data.

Systematic sampling will mean you have to work to a **system**. Take every nth pupil so that you get an even selection and also obtain your desired sample size. Taking every 30th for instance, starting with a random pupil in the first 30. (30x45 = 1350 This will give you an even spread).

In this case the data is split into natural divisions - the year groups and further into the sexes. These groups are called Strata and we could sample within these groups.

This is known as **Stratified Sampling**.

If the year 7 group contains 105 pupils, which is 7% of the population let 7% of the sample be from year 7. **Three pupils**.

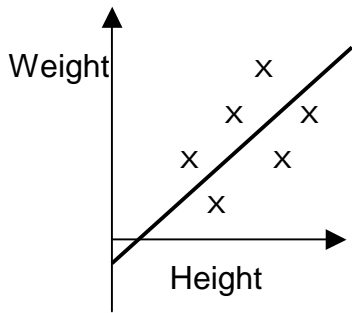
If the year 11 group contains 165 pupils - 11%, let 11% of the sample be from year 7. **Five pupils**.

Note that small numbers may arise for boys or girls in a particular year group. This could mean that results are less reliable or have greater bias.

Look out for outliers or rogue values which could influence the results.

Fitting the line of best fit

Regression - backward movement - return to an earlier stage



Draw the line so that there are an equal number of points above and below.
An improvement could be that we make the line go through a certain point.
The best point will have coordinates (mean of heights, mean of weights).
Notice the line does not go through (0,0). A zero height corresponds to a negative weight. Continuing the line downwards is called **extrapolating**.
Beware of extrapolating!

The equation of the regression line may be calculated. It is a straight line and is of the form

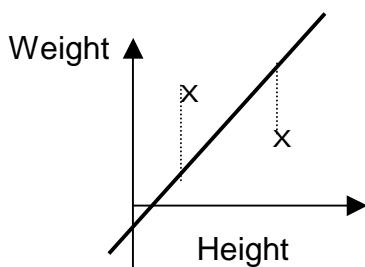
$Y = mx + c$ m is the gradient and c is the intercept on the y -axis. (The weight axis).

Is it a good fit?

We could compare the goodness of fit for lines between boys and girls say.

This can be done by looking at the two graphs and seeing that the points are generally closer to the line for one of the graphs.

We could work this closeness out numerically by averaging the distance of all the points from the line which we have constructed.



For the boys:

Measure the vertical distances of every point from the line.

Record points below the line as positive.

Find the mean of these recorded values.

Compare this mean with the mean for the girls.

The lower mean has the points closer to the line.

Also consider: A curve of best fit.

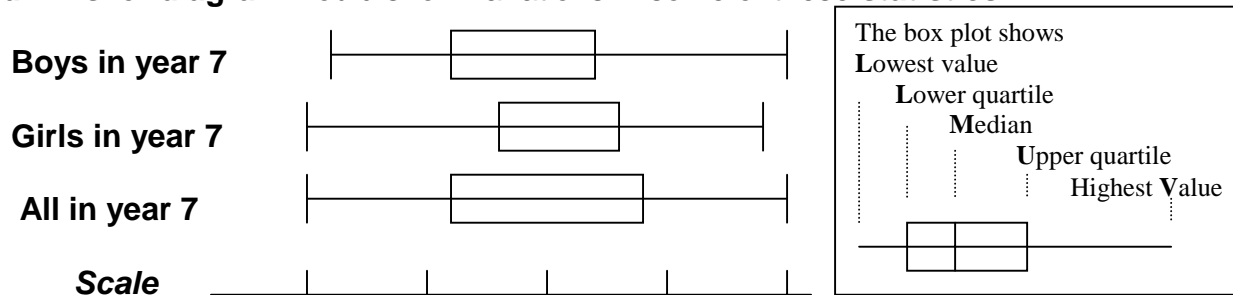
Other comparisons

We could look at heights, weights or any other numerical data and work out various measures of **location** such as the mean, median, mode or percentiles.

We could also look at measures of **central tendency** or **spread** such as the range, interquartile range, mean deviation or standard deviation.

We could look at the boys and girls in year 7 and find the above statistics for each of the genders. The quartiles and interquartile range will have to be obtained from the cumulative frequency curve.

A **box and whisker diagram** would show variations in some of these **statistics**.



We could comment on the medians with statements like "The median for the boys is greater". We could be specific with statements like "30% of the girls are taller than the upper quartile for the boys".

We could make probability statements like "The probability that a boy selected at random in year 7 is taller than the upper quartile for the girls is 0.17".

We could make similar statements from the calculations we have done. We could compare **means**, **mean deviations** and **standard deviations**.

The Standard Deviation

With a suitable set of data, we could calculate the **mean**.

The **range, a measure of spread**, is the highest - lowest but does not take account of all the other values.

Measure and record the **deviations** of all values from the mean.

These are calculated by subtracting the mean from each value in turn

The mean of these deviations will always be zero.

However, if we ignore the signs and then find the mean of the deviations we get the **mean deviation**: A measure of spread.

Another way of removing the zero result caused by negative values is to square the deviations. (Remember a square number is always positive?).

Averaging these squares gives the **variance**: Another measure of spread.

But if the data were the lengths of pencils in cm. The variance would be

Given in cm^2 . To get back to the original units take the

square root of the variance. This is the **standard deviation**.

5, 7, 7, 8, 10, 11

Mean = 8

Range = $11 - 5 = 6$

-3, -1, -1, 0, 2, 3

$(3 + 1 + 1 + 0 + 2 + 3)/6$

Mean deviation = 1.67

$(9 + 1 + 1 + 0 + 4 + 9)/6$

Variance = 4

Standard deviation = 2

Check your working: As a general rule most of the observations should lie

within two standard deviations on either side of the mean. $8 - 4 = 4$, $8 + 4 = 12$

For more detailed coverage of [Standard Deviation](#) click here.