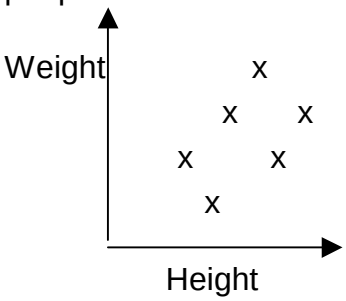
Analysing the relationship between two variables

Two variables such as **height** and **weight** could be recorded for a number of people. 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

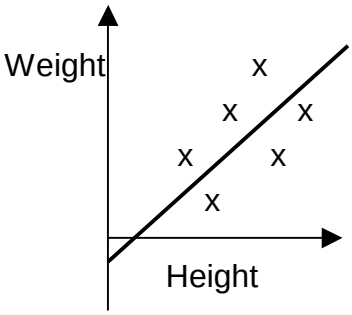
We could assume that **"as the height increases so does the weight"**.

To prove our hypothesis we could look at the whole **population.**

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

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!

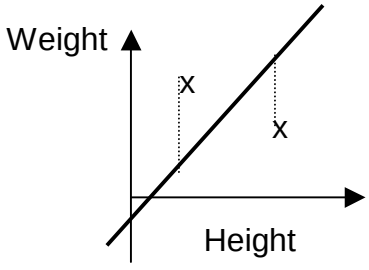
The equation of the regression line may be calculated. It is a straight line and is of the form

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

The line may be calculated and drawn on a graph using EXCEL.

Is it a good fit?

We could obtain a measure of how near the points are to the line based on the average vertical distance of the points from the line.



Measure the vertical distances of every point from the line.

Record points below the line as positive.

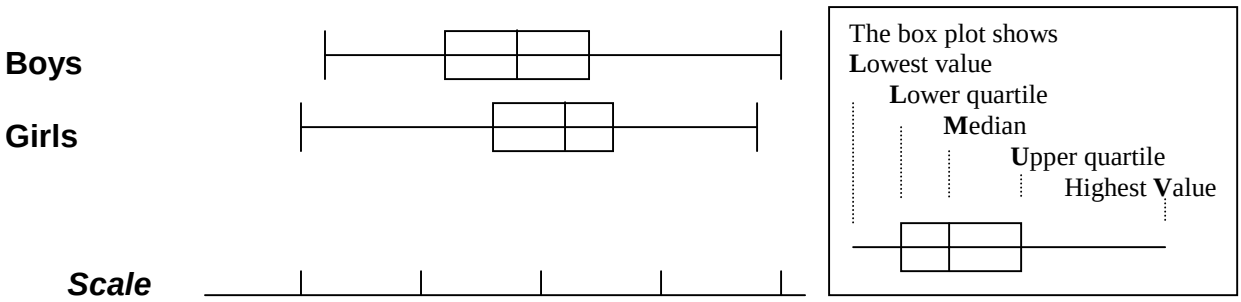
Find the **mean** of these recorded values.

The smaller this **mean**, the closer the points are to the line.

We could compare the fit for weights of boys and girls.

This mean of the vertical distances may also be calculated using EXCEL.

A **box and whisker diagram** shows key variations in a selection of data.
It is also known as a box plot and shows the Quartiles, the Range and Inter-Quartile Range.



We could comment on the medians with statements like "The median for the girls is greater".

We could be specific with statements like " 30% of the girls are taller than the upper quartile for the boys".

We could even make probability statements like "The probability that a boy selected at random, 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.