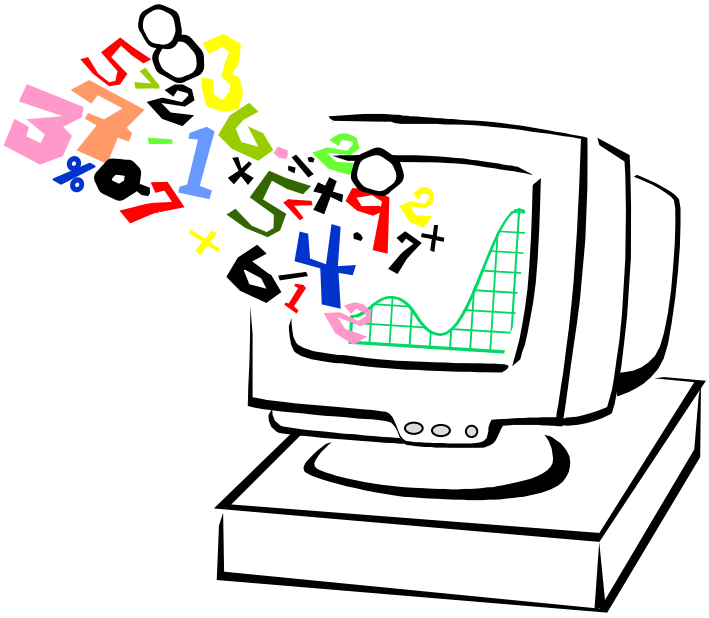
*Sheffield Hallam University*

Faculty of Health and Wellbeing
Professional Development 1
*Quantitative Analysis*



# Glossary

# Using the Glossary

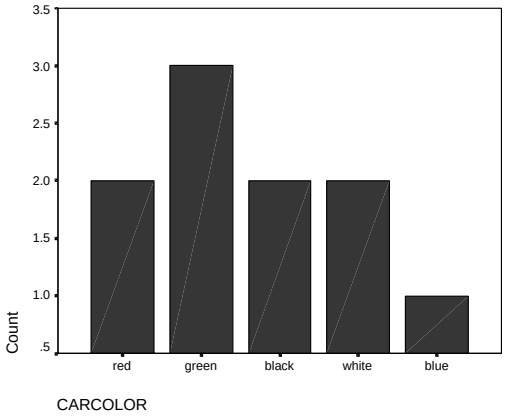
*This does not set out to tell you everything about the topics listed. Nor does it require you to learn and understand everything in it! It is hoped that what is included will help you to make sense of the concepts you are meet in your course. It should also be useful for reference when you read articles.*

*You will be directed to read certain parts as you work through the course. You will probably want to read, do an activity, and then read again with more understanding. It would be useful to skim through it all before you start, to get an idea of what you already know and what you are hoping to understand better by the end of this course.*

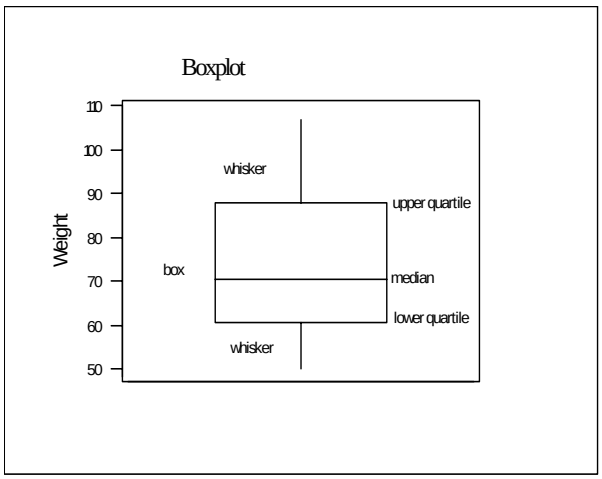
# Barchart:

Similar to a Histogram but the bars don't touch each other and the x-axis usually does not have a continuous scale.

The example shows a bar chart of the colour of car owned by ten people. (Note how the false origin gives a first impression of even less blue cars.)




# Box-plot: (also known as box and whisker plot)



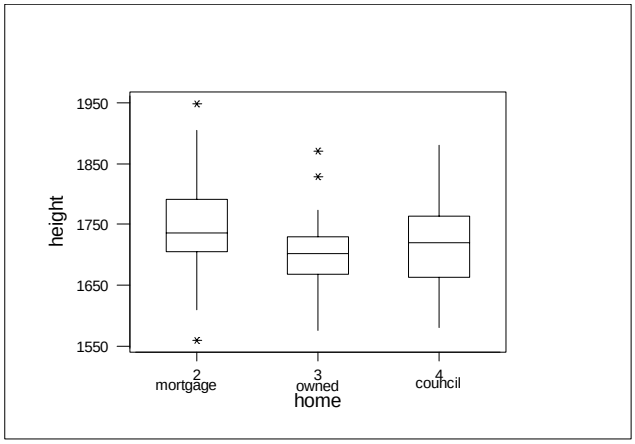

A Boxplot divides the data into quarters. The middle line shows the median (the value that divides the data in half), the box shows the range of the two middle quarters, and the whiskers show the range of the rest of the data. The values at the ends of the box are called the quartiles, (SPSS refers to these as the $25^{th}$ and $75^{th}$ percentiles) The distance between them is called the interquartile range (IQR).

The more sophisticated version (which SPSS uses) marks outliers with circles, counting anything more than one and a half times the interquartile range away

from the quartiles as an outlier, those over three times the interquartile range away from the quartiles are called extremes and marked with asterisks. The length of the box is equal to the interquartile range (IQR).

Boxplots are most often used for comparing two or more sets of data. They allow you to compare level (the median), spread (the interquartile range) at a glance, as well as showing the minimum and maximum.

The graph on the right compares the heights of men with different kinds of housing. You can see at a glance that the men who own their own houses tend to be smaller, and that there is less variation among them than among those with mortgages or in council housing. You can also see that the tallest and the smallest subjects both have mortgages.




# Correlation:

A measure of the relationship between two paired sets of data. This can be seen by eye from a scattergram.

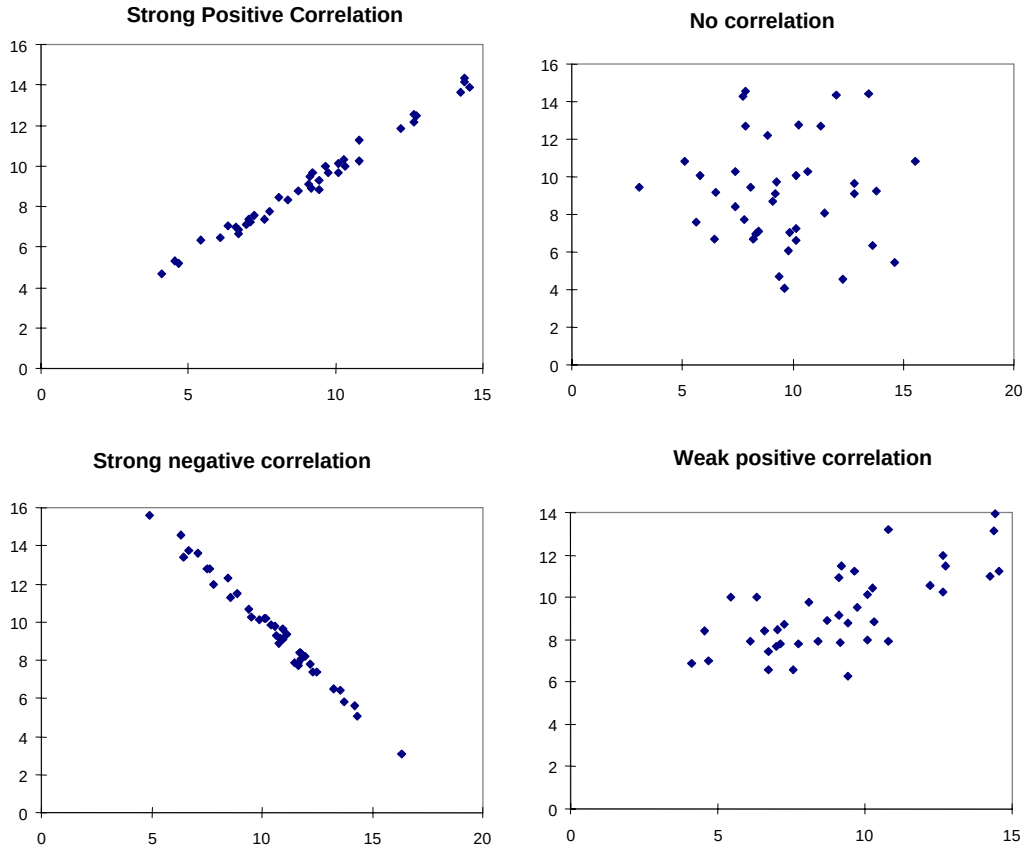
**Strong positive correlation**: The points cluster about a line that slopes upwards from bottom left to top right. Large values of one variable tend to be associated with large values of the other. *Example: Height and shoe-size exhibit a high positive correlation. Tall people tend to wear large shoes and small people tend to wear small shoes.*

**No Correlation:** The points are spread out in a way that doesn't seem to slope up or down from left to right. *Example: The number of visits to a doctor in the last six months is unlikely to be correlated with shoe-size. People with small shoes do not tend to visit the doctor more or less than people with large shoes.*

**Strong negative correlation**: The points cluster about a line that slopes downward from left to right. Large values of one variable tend to be associated with small values of the other. *Example: Percentage of patients on waiting list treated in less than 6 months and percentage of patients on waiting list treated after more than 6 months. Regions where the first is small the second will be large and vice-versa.*

**Weak positive or negative correlation:** A definite slope can be seen in the pattern of the points, but they are not so close to the line, making a shape more like an ellipse.

**Non-linear correlation:** The points cluster about a curve, not a line.




The correlation coefficient (Pearson's product-moment correlation coefficient) is a way of assigning a number to these situations. It is 1 for perfect positive correlation (all the points exactly on a line sloping up from bottom left to top right), 0 for no correlation and -1 for perfect negative correlation (all the points exactly on a line sloping down from top left to bottom right). It takes in-between values for in-between situations.

It should be noted that a high correlation coefficient on a small sample may not indicate real correlation in the background population, and that a fairly low correlation coefficient on a large sample may still indicate background correlation.

There is another correlation coefficient, known as Spearman's correlation coefficient. It is similar to Pearson's but calculated slightly differently, and less affected by extreme values. It is used in tests for correlation in circumstances where Pearson's cannot be used.

# Data (Continuous and discrete data)

A set of data is said to be **continuous** when the measurements could have any value in a range. E.g., the height of people is continuous, a person could be1812mm tall and another could be 1811.9mm tall. You may meet someone who is 1812.197347mm tall (although I doubt we could measure a human that accurately). Any number within the range of heights is possible.

A set of data is said to be **discrete** if the observations or measurements can only take on discrete values. E.g., data gathered by counting are discrete such as the number of wheels on a vehicle, the number children in a household.

Some variables might appear discrete but are actually continuous, for example scores on a patient satisfaction survey form. E.g., a patient may be asked to evaluate the quality of care received in a hospital by choosing 1, 2, 3, 4, or 5 on a feedback form. The quality of care is a continuous variable as it could take on any value between 1 and 5, however it is very difficult to discriminate between observations any more accurately when recording this observation.

Time to complete an assignment would be continuous but the number of assignments completed would be discrete.

**More examples** - try to classify each before reading the *next lin*e…

The heights of students in a seminar

*Height is continuous. For example, a student could be 162.3cm tall or any number in a range, i.e. the range of human height.*

The numbers of matches in a box.

*The number of matches is discrete. It may be 1,2,3,4…500,501… but not 2.54 or 56.79*

The times taken for a person to run 100m.

*Time is continuous. For example, an athlete may run 100m in 10.4 seconds I may take slightly longer!*

# Data (Different types)

**Nominal Data**: These are data which give classes which have no real connection with numbers and can't be ordered meaningfully.

*Examples: Male or female, Town of residence.*

**Ordinal Data**: These are data that can be put in an order, but don't have a numerical meaning beyond the order. So for instance, a distance of 2 between two numbers would not be meaningfully equivalent to a distance of 2 between two others.

*Examples: Questionnaire responses coded: 1 Strongly disagree, 2 disagree, 3 indifferent, 4 agree, 5 strongly agree.*
*Level of pain felt in joint rated on a scale from 0 (comfortable) to 10 (extremely painful).*
*Social class coded by number.*

**Interval Data**: These are numerical data where the distances between numbers have meaning, but the zero has no real meaning. With interval data it is not meaningful to say than one measurement is twice another, and might not still be true if the units were changed.

*Examples: Temperature (Centigrade), Year, adult shoe size (In all examples the zero point has been chosen conventionally, as the freezing point of water or the year of Christ's birth, or to make 1 smallest size of shoes adults were expected to wear.*

If my shoe size is twice yours in British sizes, this will not also be true in Continental sizes.

**Ratio Data:** These are data that are numerical data where the distances between data and the zero point have real meaning. With such data it is meaningful to say that one value is twice as much as another, and this would still be true if the units were changed.

*Examples: Heights, Weights, Salaries, Ages.*

Note that if someone is twice as tall as someone else in inches, this will still be true in centimetres.

**Percentage Data**: Data expressed as percentages.

*Example: Percentage of patients on waiting list operated on within 5 months.*

# Decimals, Fractions and Percentages

It is useful to be able to convert between these. If you are not happy with converting between fractions, decimals and percentages it is worth reminding yourself of the following and working out a few for yourself, so you don't panic if you meet something in an unfamiliar form.

Percentages to decimals: divide by 100. *e.g. 7% = 0.07* or 50% = 0.5

Decimals to percentages: multiply by 100. *e.g. 0.003 = 0.3% or 0.25 = 25%*

Fractions to decimals: divide the top by the bottom. *e.g. 3/8 = 3 ÷ 8 = 0.375*

Decimals to fractions: Put the decimal places over 10, 100, or 1000 etc. depending on how many there are. *e.g. 0.3 = 3/10, 0.04 = 4/100, 0.007= 7/1000*. You can then often simplify these by dividing the top and the bottom by a common factor, or using a calculator that does this for you: *e.g. 4/100 =1/25*.

Percentages to Fractions: If it is a simple whole number put 100 underneath it and simplify if necessary. Otherwise turn it into a decimal first.
*e.g. 5% = 5/100 = 1/20, 3.7% = 0.037 = 37/1000*

Fractions to Percentages: If there's 100 on the bottom, leave it off. Otherwise turn it into a decimal first. *e.g. 3/100 = 3%, 7/200 = 7 ÷ 200 = 0.035 = 3.5%*

# Dependent and Independent Variables:

See explanatory and response variables.

# Descriptive Statistics:

A general term for ways of describing a sample without attempting to draw conclusions about the background population. The mean, median, standard deviation and inter-quartile range are examples of descriptive statistics, as are graphs.

# Explanatory and Response Variables:

In a situation where we have a hypothesis that changes in one variable explain changes in another, we call the first the explanatory variable and the second the response variable (because it responds to changes in the first). A scattergram should always have the explanatory variable on the x-axis and the response variable on the y-axis.

> *Example: the hypothesis is that your heart rate increases the longer you exercise. You control the time of exercise by taking*

*measurements of heart rate after 0, 5, 10, 15 minutes etc. Time is the explanatory variable and heart rate is the response variable.*

A hypothesis that changes in one variable explain changes in another is best tested in a situation where the explanatory variable can be controlled, as in the above example.

In medical statistics, situations where one variable is controlled can be difficult to set up ethically. *(How would patients react in your discipline if they were told the length of their treatment would be decided at random as part of an experiment?)*

This means we often cannot choose people at random to give different treatments, but must use the treatments they were given for other reasons. This may mean that the explanation for the response variable comes not from the different treatments, but from other different factors that determined the treatments.

*Example: it was argued for a long time that heavy smokers did not die of lung cancer because they were heavy smokers, but because of other lifestyle factors which drove them to become heavy smokers.*

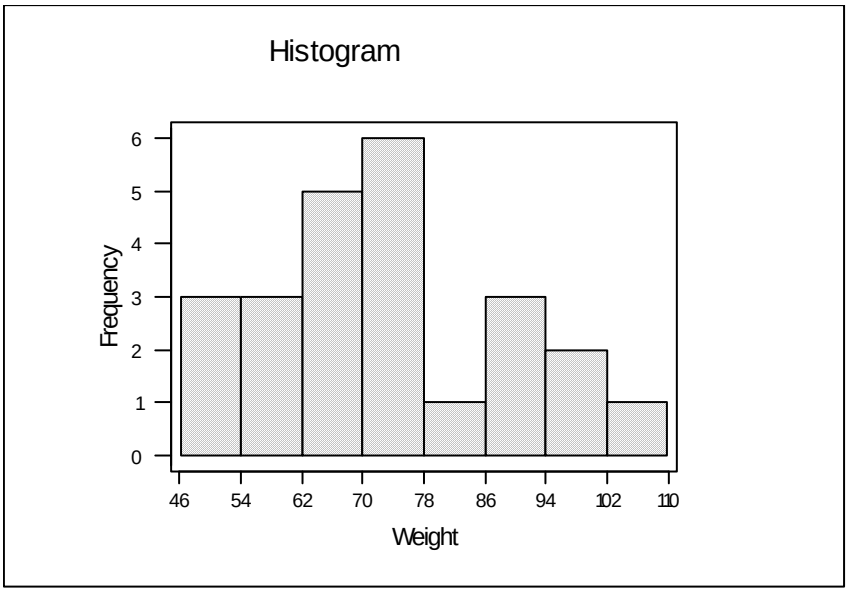
There are many situations where variables are correlated but neither is explanatory.

*Example: Areas where more households have two cars also report less deaths from lung cancer. Both variables are at least partly explained by the variable money.*

In situations where the explanatory variable is controlled experimentally it is often known as the independent variable, and the response variable as the dependent variable (as you can decide independently what the independent one will be, and the other depends on it).

# Histogram:

A kind of barchart where each bar represents the frequency of a group of data between certain values. The bars touch each other and the x-axis has a continuous scale. (Not the case in other types of bar chart, where the data does not need to be continuous.)




Histograms are usually used to examine the distribution of data: whether they are evenly spread out along the range, or bunched together more at some points. In particular, a histogram is one way of checking whether data are roughly normally distributed.

# Inter-quartile Range:

A measure of spread or variability, similar to the standard deviation**.** It is most often used to compare the variability of different samples.

It is the difference between the lower quartile and the upper quartile. These are the values that a quarter of the data lies below, and that three quarters of the data lie below, so the inter-quartile range is the range of the middle half of the data.

> *Example: A group of 12 patients has ages 18, 18, 19, 19, 19, 20, 21, 23, 30, 33, 45, 81. The lower quartile is 19 and the upper quartile is 31.5. The interquartile range is 12.5. (31.5 - 19 = 12.5)*
>
> *Another group of 12 patients has ages 18, 19, 19, 19, 19, 19, 20, 21, 21, 22, 22, 85. The lower quartile is 19 and the upper quartile is 21.5. The interquartile range is 2.5. The first group has more variability in age.*

Box-plots show the quartiles.

SPSS will calculate the quartiles and the inter-quartile range can be calculated easily from these by subtracting the lower quartile from the upper one.
(There is some disagreement in different books about the exact method of calculating quartiles - all different methods come out pretty close and we are not concerned here with the details.)

# Mean (Arithmetic mean):

A measure of level or central tendency, the mean gives a number somewhere in the middle of your data set. The Mean is often referred to as the average, but this can cause confusion as the Median and the Mode are also kinds of averages. The mean is calculated by adding up all the data and dividing by how many there are. SPSS will do this for you on the computer. Most scientific calculators will also give you means directly.

> *Example: A sample of 5 patients have ages 18, 23, 20, 18, 81. The mean is (18+23+20+18+81) ÷ 5 = 32. Note that this mean is considerably larger than 4 of the ages in the set. If the 81 had in fact been mistyped for 18 your result would be seriously affected by this.*

The mean has the advantage over the median that it takes into account all the data, and the disadvantage that very large or very small values can have a distorting effect on it.

# Mean (Geometric mean):

Another measure of level or central tendency but much more difficult to calculate! Rather than adding the numbers together and dividing by the number of numbers, the numbers are multiplied together and for "N" numbers the $N^{th}$ route of the result is taken. When people refer to the mean they usually mean the Arithmetic mean, so don't worry about the geometric mean. I include it here mainly for completeness.

# Median:

Another measure of level or central tendency. The median is found by ranking the data set in order and taking the middle one (or the mean of the two middle ones if there are two).

> *Example: A sample of 5 patients have ages 18, 23, 20, 18, 81. In order this is 18, 18, 20, 23, 81. The median is 20, the middle value. If a patient's age lies below the median they are in the bottom half of the set, and if above the median they are in the top half.*

The median has the advantage over the mean that it is often easier to see by eye for very small data sets, and is not unduly affected by extreme values. It can be calculated on SPSS and some calculators. It is useful when you want to know whether a particular result lies in the top or bottom half of a data set.

Box-plots show the median.

In a symmetrical distribution, the mean and the median will be close. Differences between the mean and median indicate asymmetry.

# Mode:

The most frequent data value. It is often the easiest to pick out by eye.

> *Example: A sample of 5 patients have ages 18, 23, 20, 18, 81. The mode is 18, since this age occurs most often.*

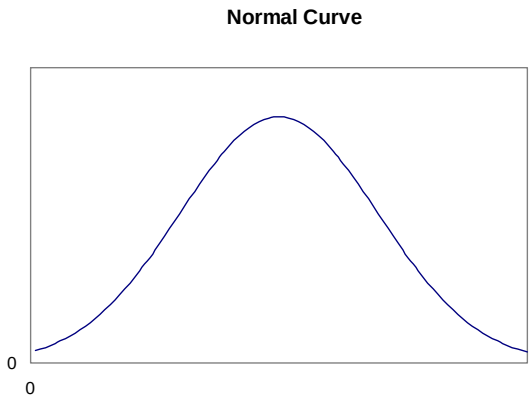
In a roughly normal distribution the mode will be close to the mean and the median.

It is possible for a data set to have several modes. The presence of several modes in a large dataset can indicate that different populations have been combined.

# Normal Distribution:

The name of a specific distribution with a lot of data values near the mean, and gradually less further away, symmetrically on both sides. A lot of biological data fit this pattern closely.

The histogram for a large number of normal data has a bell-shaped curve.




# Outlier:

A data value, which is very big or very small, compared with the others. Sometimes these are due to mistakes in entering the data and should always be checked.

Outliers which are not mistakes can occur. It is worth examining your data carefully and trying to explain why certain items stand out.

There are different rules for deciding exactly what to count as an outlier.

In SPSS a circle on a boxplot is used to mark outliers with values between 1.5 and 3 box lengths from the upper or lower edge of the box. (The box length is the interquartile range.)

In SPSS an asterisk on a boxplot represents an extreme outlier (just called an extreme in SPSS documentation but I feel the term extreme outlier is more helpful) which is a value more than 3 times the interquartile range from a quartile.

# Paired Data:

Data are paired if the entries in each row are connected with each other.

*Examples:*
*Paired:*

- *the ages and weights of a group of gymnasts*
- *the weights of a group of gymnasts before and after a training session*

*Non-paired:*

- *the weights of a group of gymnasts and a group of non-gymnasts*
- *the changes in weight of two groups of gymnasts given different kinds of training session*

If you are not sure whether two columns of data are paired or not, consider whether rearranging the order of one of the columns would affect your data. If it would, they are paired.

Paired data often occur in 'before and after' situations. They are also known as 'related samples'. Non-paired data can also be referred to as 'independent samples'.

Scatterplots (also called scattergrams) are only meaningful for paired data.

# Pie chart:

Pie charts, are used to show proportion, e.g. the number of votes cast for each party in an election. The pie should add up to 100% of the observed data. The size of each slice is proportional the percentage of the data it represents.

# Population:

The background group that we are using the sample to find out about.

*Example: A group of 20 patients with anxiety problems are used to draw conclusions about how any patients with anxiety problems would respond to treatment. The population could be: patients in Sheffield with similar problems, patients in England, patients all over the world, patients from similar ethnic groups etc.*

Conclusions may be more or less valid depending on how wide the population they are supposed to apply to is, and how representative of that population the sample is. Strictly, a sample should be drawn at random from its population.

# Range:

The difference between the smallest and largest value in a data set.

It is a measure of spread or variability, but only depends on the two extreme values, and does not tell anything about how spread out the rest are.

It can be distorted by one extreme value.

> *Example: a group of patients are aged 18, 20, 23, 18, 81. The range is 63. The 81 year old has a huge effect on this: if it were a mis-typing for 18 the result would be very distorted.*

It is useful as a very quick measure of variability, but the inter-quartile range or the standard deviation are to be preferred for more precise comparisons between different data sets.

# Sample:

The group of people, (or things, or places,) that the data have been collected from. In most situations it is important to pick a representative sample, which is not biased *e.g. mainly women, mainly from particular age or income bands or with particular educational qualifications*. There is a range of methods for doing this.

# Scatterplots (Also known as x-y plots and Scattergrams):

A graph used to show how paired data are related.

Each point represents a pair of data values, one given by its x co-ordinate and the other by the y co-ordinate. They are used to look for correlation.

They can also be used to look for increases or decreases after a treatment, by plotting before and after values and seeing whether most of the points lie above or below the y = x line.

See the graphs used to illustrate correlation for examples of scattergrams.

# Standard Deviation:

A measure of the spread or variability of a data set.

The larger the standard deviation, the more spread out about the mean the data are.

Like the mean, the standard deviation takes all values into account and can be greatly influenced by extreme values. The Inter Quartile Range is less effected.

You can find how to calculate it in any standard statistics book but you do not need to, as SPSS will calculate it for you. Most scientific calculators will also calculate it from the raw data if you do not have access to a computer.

> *Example: Two groups of 5 patients have the following ages: Group A: 18, 24, 30, 36, 42, Group B: 18, 19, 20, 38, 55, . Both groups have the same mean, 30. The standard deviations are 8.5 for Group A and 14.5 for Group B, showing the ages in Group B are more spread out from the mean.*

# Variance:

The square of the standard deviation.

It is used a lot in statistical calculations, but you won't need it to use and interpret statistics. The standard deviation is the square root of the Variance.

# X and Y axes and co-ordinates:

The x-axis is the horizontal line along the bottom of a graph and the y-axis is the vertical line up the side, (except where negative values are involved, when the axes will be in the middle of the graph). Any point on a graph has an x co-ordinate, which is the number on the x-axis level with it, and a y co-ordinate, which is the number on the y-axis level with it.
The point where both co-ordinates are zero is called the origin.
The diagonal line which goes through all the points whose x and y co-ordinates are the same is called the line y = x.