4 Descriptive statistics

What are descriptive statistics?

Even a quite small experiment can generate large amounts of data; columns of times, or numbers of errors, or some other scores. It can then be difficult to see what has been going on. This calls for ways in which you can summarize the data. Descriptive statistics do this – they are, in fact, sometimes referred to as 'summary statistics'.

This chapter concentrates on two aspects of a set of data which are commonly summarized. These are covered by statistics which describe the **most typical value** (some kind of **average**); and how much **variability** there is about this central value. There is then discussion of a third statistic which describes the relationship between two sets of data (known as **correlation**).

While these descriptive statistics are extremely useful, do not neglect the opportunities that graphs of various kinds present in displaying and summarizing data. We already came across this in the previous chapter where the **histogram** of Figure 4 provides a much more vivid representation of the probabilities of different numbers of +s than does Table 6. It is worth noting that, when displaying a variable in the form of categories (e.g. 'female' and 'male'; or different ethnic origins) it is the convention to show these as separate, non-touching bars as in Figure 6.

Such graphs (sometimes referred to as **bar charts** or **bar graphs**) emphasize that there are no intermediate values and that the ordering along the axis (males on the left or right) is arbitrary.

For similar reasons the use of standard **line graphs** should be restricted to situations where the variable along the horizontal axis

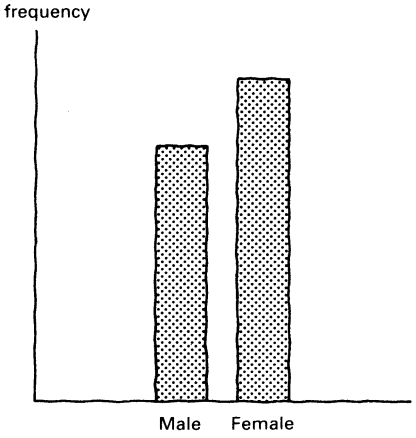

Figure 6 Histogram (bar chart) with categorical variables

is continuous (e.g. age or height). Only then are intermediate values between the points on the graph meaningful.

Measures of central tendency

1 The mean

This is a very commonly used measure of *most typical value.* You probably already know it as the average, obtained by adding all the scores together and then dividing by the number of scores.

It can be used to provide an introduction to some of the symbols widely used in statistics. Scores in general are commonly represented by the symbol X; the first score by X_1; the second by X_2 and so on. If there is a total of N scores, then the last of these is represented by X_N. The mean itself is given a special symbol $\bar{X}$, usually referred to as 'X bar'.

$$\bar{X} = \frac{\text{total of all scores}}{\text{total number of scores}} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}.$$

This can be simplified by the use of the 'summation' instruction 'Σ'. Placed in front of a symbol, such as 'X', it means 'add all the Xs together'.

$$\bar{X} = \frac{\Sigma X}{N}$$

(Note that ΣX does *not* mean 'multiply Σ by X'.)

For example, if the scores are

7, 3, 11, 12, 9, 14

then $\Sigma X = 7 + 3 + 11 + 12 + 9 + 14 = 56$

and $N = 6$

so

$$\bar{X} = \frac{\Sigma X}{N} = \frac{56}{6} = 9{\cdot}3,$$

i.e. the mean is 9·3.

Strictly speaking, the proper label for an average calculated in this way is the 'arithmetic mean'. This is to distinguish it from other kinds of mean (the 'geometric' and 'harmonic' means) which are used for special purposes. However, the arithmetic mean is much more commonly used and will be the one assumed if you simply refer to the 'mean'.

2 The median

The median is the central value in a set of scores. It is obtained by arranging the scores in order of size. With an odd number of scores the median is simply the score which then has equal numbers of scores above and below it. With an even number of scores, it is the average of the two central scores. (There are more complicated formulae for calculating the median with an even number of scores, but the result seldom differs appreciably from simply taking the average of the two central scores.) If there is a cluster of scores around the centre, all having the same value, then the simplest procedure is to regard that value as being the median.

As an example, if the scores are:

14, 9, 17, 21, 7, 18, 16, 22

then, rearranging these scores in increasing order of size we get:

7, 9, 14, 16, 17, 18, 21, 22.

As there are eight scores, the median is the average of the fourth and fifth scores:

$$\frac{16 + 17}{2} = 16{\cdot}5$$

i.e. the median is 16·5.

3 The mode

The mode is the value which occurs most frequently in a set of scores. It usually only makes sense to use the mode as a measure of central tendency when you have a large set of scores. Even then it may be necessary to group scores together (i.e. to put together all scores in a certain range, say from 21 to 25 cms, 26 to 30 cms, etc.).

If a histogram is plotted, the highest frequency is given by the longest bar in the histogram and hence gives you the mode. Figure 7 shows an example where grouping has taken place. When dealing with scores which are ordered in this way, it is possible for there to be a second major peak in the distribution of scores displayed in a histogram. This is called a **bimodal** distribution and is best dealt with by displaying the histogram rather than by quoting the mode.

Mean, median and mode compared

The mean is the statistic most commonly used as a measure of central tendency. One reason for this is its sensitivity; in the sense that if any one of the scores in a set of scores changes, then the mean will change. In contrast, both the median and the mode may well be unaffected by changing the value of several scores. Try this out for yourself.

While this sensitivity is often an advantage it can be a disadvantage. Suppose the following scores represent times in seconds to solve some anagrams:

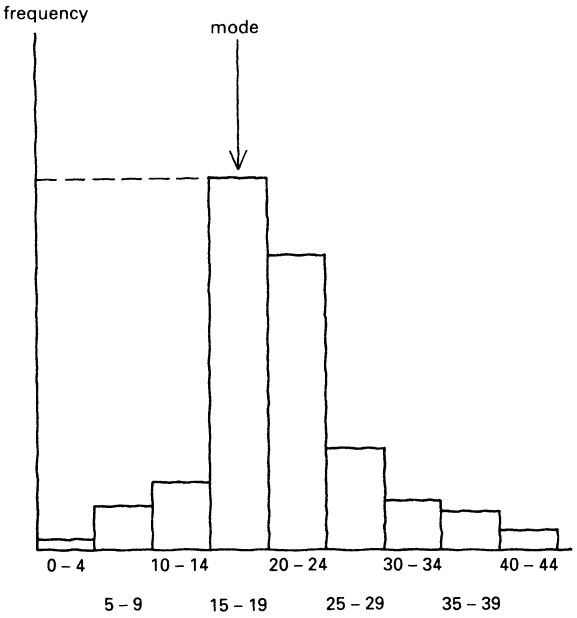

Figure 7 Mode derived from histogram

1, 2, 2, 3, 4, 5, 5, 6, 224

the mean works out at 28. However, this mean is a very strange 'most typical value'. What we have in the set of scores is a quite tightly clustered group well below ten seconds, and a 'rogue' score (the technical term for this is an **outlier**) where, perhaps, the person involved had some kind of block in solving the anagram. 28 seconds is not typical of either. Here, the median (4 seconds) would be the most appropriate central tendency statistic to use, simply because it is insensitive to the values of extreme scores.

Note that the median (and the mode) can still be calculated when some extreme values are unknown; say when a participant has such a block that they find an anagram impossible to solve and all you can record is 'over 5 minutes'.

There are some kinds of data where it is not possible to calculate

a mean. An example is where your data is in the form of rankings or orderings (i.e. you know which is first, highest, longest, etc.; which is second and so on, but don't know any actual scores). It is still feasible to work out a median though. If you have the kind of data displayed in Figure 7, where all you know are frequencies and there is no appropriate way of ordering the data, then neither mean nor median is feasible and you are left with the mode as the only possibility.

Measures of dispersion

Sets of scores with the same mean may be very different from each other. Consider the set:

17, 32, 34, 58, 69, 70, 98, 142

and a second set of:

61, 62, 64, 65, 65, 66, 68, 69.

Both have means of 65, but the **dispersion** (otherwise known as **variability** or **spread**) of the second set is much smaller than that of the first set. Several statistics have been devised to measure this aspect of a set of data.

1 Range

The range is the difference between the highest and lowest scores. It is, therefore, very easy to compute. As an example, take the following scores arranged in order of size:

19, 21, 22, 22, 25, 27, 28, 42.

The range is highest minus lowest,

i.e. **range** = 42 − 19 = 23.

The main disadvantage of the range as a measure of dispersion is that it is just based on these two extreme scores. Such scores may be suspect and it may be unwise to give them undue weight. For example, an abnormally long time to respond in some task may be due to the participant day-dreaming or otherwise not attending rather than to the difficulty of the task.

2 Semi-interquartile range

This is a more sophisticated type of range statistic. If the scores are arranged in ascending order of size, the point that cuts off the lowest quarter of the scores is called the **first quartile** (Q_1). The point that cuts off the lowest three quarters of the scores is called the **third quartile** (Q_3). If, for instance, there are 24 scores, the first quartile (Q_1) occurs between the 6th and 7th scores (take the average of 6th and 7th scores, as for the median). The third quartile (Q_3) occurs between the 18th and 19th scores (again, take their average).

The interquartile range is the difference between the third and first quartiles. As the name suggests, the semi-interquartile range is half of this,

i.e. **semi-interquartile range** $= \dfrac{Q_3 - Q_1}{2}$.

This measure of range is commonly used when the median is used as a measure of central tendency. (Note that the median, the central point in a set of scores, has half of the scores below it and can also be referred to as the **second quartile**.)

The relative sizes of the differences $(Q_3 - Q_2)$ and $(Q_2 - Q_1)$ provides a useful measure of the **skewness** (or lack of symmetry) of the distribution of a set of scores. Figure 8 shows three histograms. One has a longer 'tail' to the left (lower scores) than to the right (higher scores). This is called a **negative skew**. A second one is symmetrical. The third has a longer tail to the right (higher scores) than to the left (lower scores). This is called a **positive skew**.

In terms of the quartiles:

There is a negative skew if $(Q_2 - Q_1) > (Q_3 - Q_2)$,
and positive skew if $(Q_2 - Q_1) < (Q_3 - Q_2)$.

The advantage of the semi-interquartile range over the range as a measure of dispersion is that it is not simply dependent on the two extreme values.

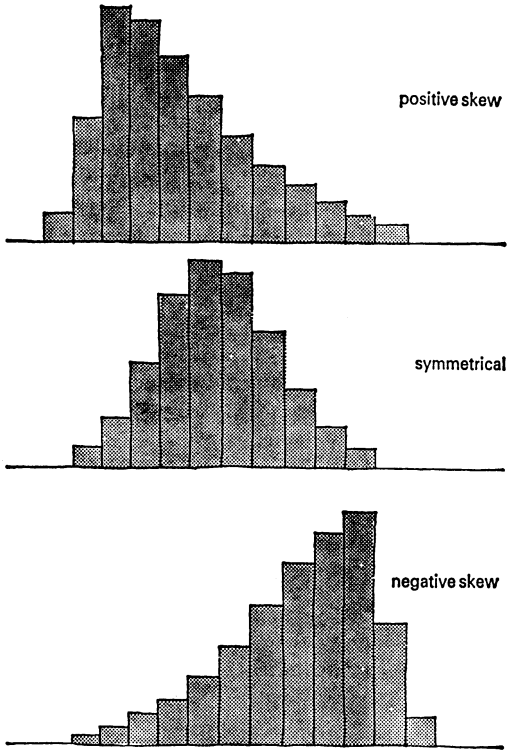

Figure 8 Positive and negative skew

3 Mean deviation

The **deviation** (x) of a score is the difference of that score from the mean. In symbols, if X is a score, and $\bar{X}$ the mean score, then the deviation x is given by

$$x = X - \bar{X}.$$

On first thoughts, it might appear very sensible to use the average of such deviations as a measure of dispersion. However, if you do this, taking note of the fact that some deviations will be positive

and some negative, you will find that the average always comes out at zero! (Try it.)

One way of rescuing this situation is to ignore the signs of the deviations. This is what is done when using mean deviation as a measure of dispersion. Hence, **mean deviation** ($\bar{x}$) is given by the formula

$$\bar{x} = \frac{\Sigma|X - \bar{X}|}{N}$$

where N is the number of scores; Σ is the instruction 'take the sum of'; and $|X - \bar{X}|$ means 'take the absolute value of the difference between X and $\bar{X}$' (in other words, always call the difference positive, i.e. take the smaller from the larger).

Compared with the two previous range-based statistics, the mean deviation has the advantage that it is based on all the scores. It is not widely used though, largely because it has been supplanted by other measures of dispersion based on deviations, which are considered below.

4 Variance

An alternative tactic to get over the problem that deviations always add up to zero when sign is taken into account, is to make use of squared deviations. These are always positive (as 'minus times minus is plus').

The mean of the squared deviation is a commonly used statistic and is called the **variance**. From our previous discussions you would expect the formula for variance to be

$$\text{Variance} = \frac{\Sigma(X - \bar{X})^2}{N}$$

where X is the individual score
$\bar{X}$ is the mean score
$X - \bar{X}$ is the deviation
$(X - \bar{X})^2$ is the squared deviation
Σ is the instruction 'take the sum of'
hence $\Sigma(X - \bar{X})^2$ means 'take the sum of the squared deviations'

and N is the total number of scores.

This formula does give the variance of the actual set of scores you have in front of you. This is appropriate when you have a sample of scores and simply wish to describe various aspects of that sample. Or if you have measures from the whole of a population of some kind.

However, as has been stressed previously, the situation is more usually that we have a sample and want to make estimates about the state of affairs in the population from which the sample has been drawn. In these circumstances it can be shown (magic words!) that an unbiased estimate of the variance in the population is obtained by use of a slightly different formula:

$$\text{Variance} = \frac{\Sigma(X - \bar{X})^2}{N - 1}.$$

The $N - 1$ in the formula, which is substituted for the N in the previous formula, is referred to as the **degrees of freedom**. As the name suggests it is the number of deviations from the mean which are free to vary. This is one less than N because the final deviation is fixed by the need for the overall average deviation to be zero. This is by no means an adequate explanation for the use of $N - 1$ in the formula, but there is at least an intuitive rightness to the notion that our estimate of dispersion or variability should be divided by the number of things free to vary.

The effect of this change to the formula is to increase the estimate of variance. However, this makes little difference unless very small samples are used.

5 Standard deviation

A disadvantage of the variance as a measure of dispersion is that it is in squared units as compared with the original data (e.g. seconds squared rather than seconds). A simple solution to this is to take the square root of the variance. This is known as the **standard deviation (SD)**:

$$\text{SD} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}.$$

This is by far the most commonly used of the measures of dispersion and would usually be paired with the mean as a measure of central tendency. Its popularity also arises from its links to the 'normal' distribution and various statistics associated with this distribution, as discussed in the next chapter.

If you are using a calculator (and we are now getting to statistics where it can be a substantial chore to do things by hand) a different version of the formula is marginally easier to deal with. This is

$$\text{SD} = \sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/N}{N - 1}}$$

where X is the individual score
X^2 is the individual score squared
ΣX^2 is the sum of the individual scores squared
ΣX is the sum of the individual scores
$(\Sigma X)^2$ is the square of the sum of the individual scores
and N is the total number of scores.

This formula can be derived by algebra from the previous one and hence leads to exactly the same result (assuming no arithmetical errors). You might like to check this using a small set of scores.

Just a word of warning. Take especial care not to confuse ΣX^2 – where you first square the scores and then add those squares together; and $(\Sigma X)^2$ – where you first add all the scores and then square the total.

Step-by-step procedure

Standard deviation

What to calculate	*How to calculate it*	*Usual symbols*
Step 1 total	Add all the observations together	ΣX
Step 2 mean	Divide the result of **step 1** by the number of observations	$\frac{\Sigma X}{N} = \bar{X}$
Step 3 uncorrected sum of squares	(a) Square each of the observations	X^2
	(b) Add all the squares together	ΣX^2
Step 4 correction term	(a) Go back to the total obtained in **step 1** and square it	$(\Sigma X)^2$
	(b) Divide the result of **step 4a** by the number of observations	$\frac{(\Sigma X)^2}{N}$
Step 5 corrected sum of squares	Subtract the result of **step 4b** from that of **step 3b**	$\Sigma X^2 - \frac{(\Sigma X)^2}{N}$
Step 6 variance	Divide the result of **step 5** by (number of observations − 1) NB $(N - 1)$ is often referred to as 'degrees of freedom'	$\frac{\Sigma X^2 - (\Sigma X)^2/N}{(N - 1)}$
Step 7 standard deviation	Take the square root of the result of **step 6**	$\sqrt{\frac{\Sigma X^2 - (\Sigma X)^2/N}{(N - 1)}}$

Worked example

Standard deviation

Observations		**Step 3a** $(Observations)^2$
4·5		20·25
6·0		36·00
7·4		54·76
8·2		67·24
2·1		4·41
6·5		42·25
5·4		29·16
9·3		86·49
10·8		116·64
8·0	**Step 3b uncorrected sum of squares**	64·00
Step 1 total = 68·2		521·20

Step 2 mean $= \dfrac{68{\cdot}2}{10} = 6{\cdot}82$

Step 4a $= (68{\cdot}2)^2$

Step 4b correction term $= \dfrac{(68{\cdot}2)^2}{10} = \dfrac{4651{\cdot}2}{10} = 465{\cdot}12$

Step 5 corrected sum of squares $= 521{\cdot}20 - 465{\cdot}12 = 56{\cdot}08$

Step 6 variance $= \dfrac{56{\cdot}08}{9} = 6{\cdot}23$

Step 7 standard variation $= \sqrt{6{\cdot}23} = 2{\cdot}5$

Standard scores

For purposes of comparison, say, of different individuals on the same test, or of the same person on different tests, it is often useful to transform scores into **standard scores** (otherwise known as **z-scores**). This is done by using deviations expressed in terms of standard deviation units:

$$\text{standard score } (z) = \frac{\text{deviation score } (x)}{\text{standard deviation (SD)}}$$

where $x = X - \bar{X}$.

When all scores in a set of scores are transformed into z-scores the distribution is said to be standardized. This point is returned to in the next chapter.

Correlation

Measures of central tendency and dispersion are ways of describing and summarizing sets of scores on a single variable. However, we may well have data on two or more variables and want to look at the relationship between scores on the different variables. Suppose we have both error and time scores on a particular task from a group of participants. It may be that those who are good at the task do it quickly and make few errors. Those who are poor at it take longer and make more errors. Or there might be an entirely different relationship between scores on the two variables. Those who do it quickly might make many errors. Those doing it slowly and carefully make few errors. Such relationships are known as **correlations** between scores on the two variables, i.e. they are co-relationships.

A **positive correlation** is when high scores on one variable tend to be paired with high scores on the second variable (e.g. when individuals make many errors and take a long time); and low scores on one variable tend to be paired with low scores on the other (e.g. when they make few errors and take a short time).

A **negative correlation** is when high scores on one variable tend to be paired with low scores on the second variable (e.g. when

individuals make many errors and take a short time); and low scores on one variable tend to be paired with high scores on the other (e.g. when they make few errors and take a long time).

Various **correlation coefficients** have been devised which give a numerical value for the correlation. They give values ranging from +1 for a perfect positive correlation, through zero for no correlation, to −1 for a perfect negative correlation. Intermediate values give an indication of the strength of the relationship between the two variables. We will consider how one such coefficient can be calculated later in the chapter.

Scattergrams

A simple and useful way of displaying the relationship between two variables is provided by the **scattergram** (sometimes referred to as a **scatterplot**). This involves plotting one of the variables along the horizontal dimension and the other along the vertical dimension. If the two variables are the independent variable and dependent variable in an experiment then the convention is to plot the independent variable on the horizontal axis, and the dependent variable on the vertical axis (this applies on all graphs).

However, in considering correlation, it may well be that both variables are dependent variables; or they may arise from some non-experimental situation where it is not appropriate to make the distinction between independent and dependent variables. Commonly scores arise from individuals being measured or otherwise contributing data on two variables, so that for each participant there is a pair of scores.

Suppose, for example, we obtain measures of height and weight for a class of children. To plot a scattergram each child contributes one dot, positioned according to their height and weight. Thus a subject with height 135 cm and weight 31 kg would be represented as shown in Figure 9.

For a set of subjects, the scattergram might look like Figure 10. This shows that there is some relationship between the variables. Tall children tend to be heavier. This is an example of a positive correlation, in that high scores on one of the variables (height) tend to be associated with high scores on the other variable (weight).

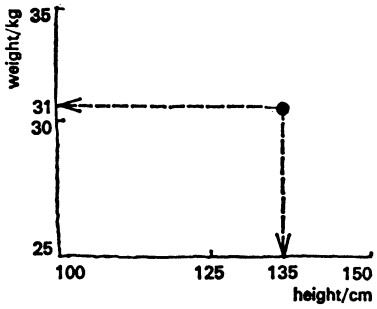

Figure 9 Representation on a scattergram of a participant with height 135 cm and weight 31 kg

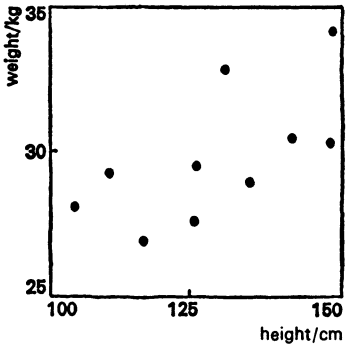

Figure 10 An example of a scattergram

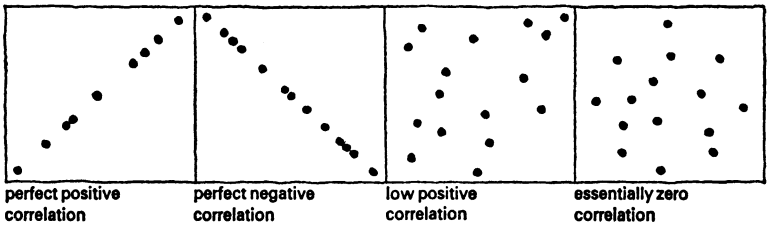

Figure 11 Scattergrams for different degrees of correlation

Examples of several different kinds of relationship are shown in Figure 11.

An alternative way of thinking about correlation is to say that a high or strong correlation enables us to make accurate predictions about an individual's score on one variable when we know their score on the other variable. Note that this applies with equal force for both positive and negative correlations. If there is a strong negative correlation, it simply means that we predict low scores for an individual on one variable if they score highly on the other variable.

Spearman's rho

Several correlation coefficients have been developed. We will cover just one of these here, Spearman's rho (ρ), otherwise known as Spearman's rank order correlation coefficient. Appendix 2 gives details of a second correlation coefficient, Pearson's r.

Spearman's rho is based on rank orders. It deals, not with the scores themselves, but with the order of these scores when they have been ranked in size. There are, of course, some situations where ranks or orderings are all we have. Say that we can measure preferences within a set of things so that one of them is placed first (given rank one), another is placed second (given rank two) and so on. Spearman's rho can also be used in situations like this where the data are in the form of ranks from the start.

Suppose we have two people ranking a set of eight politicians on some quality, say honesty. The ranks may be as follows:

Table 7 Rankings of politicians' honesty given by two persons

	Rank given to politician							
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*
person 1	1	2	3	4	5	6	7	8
person 2	1	4	2	5	8	6	3	7

We have taken the first person's ordering as the basis and shown how the second person compares – thus they agree about who is most honest, but the politician ranked second by the first person is

ranked fourth by the second person. In attempting to measure the correlation between these rankings it is clear that if they are perfectly correlated we have two possibilities. Either the rankings given by the two persons are identical (a perfect positive correlation); or one ranking is the reverse of the other. Whoever person one ranks first is ranked last by person two, and so on (a perfect negative correlation).

Spearman's rho is based upon the amount of disagreement between the two rankings. Specifically, the measure used is the sum of the squared difference in ranks (i.e. Σd^2, where d is the difference in rankings for each of the things ranked).

In the example given above, $d = 0$ for politician A, $d = -2$ for politician B, $d = 1$ for politician C, etc. and

$$\Sigma d^2 = 0^2 + (-2)^2 + 1^2 + (-1)^2 + (-3)^2 + 0^2 + 4^2 + 1^2 = 32$$

It is clear that Σd^2 will be a minimum, in fact zero, when the two rankings are identical. Similarly Σd^2 will be a maximum when one rank order is the exact reverse of the other. Hence the equation

$$\text{Spearman's rho } (\rho) = 1 - \frac{2\,\Sigma d^2}{\text{maximum value of } \Sigma d^2}$$

gives a correlation coefficient of $+1$ when there is no disagreement ($\Sigma d^2 = 0$) and a correlation coefficient of -1 where the disagreement is a maximum. You can check this by substituting Σd^2 for 'maximum value of Σd^2' into the equation for Spearman's rho.

It is not unreasonable to think of a total lack of correlation as being half-way between these two extremes of perfect agreement and perfect disagreement. If we take the value of Σd^2 as half of its maximum value, you find that this produces a value of zero for Spearman's rho. If Σd^2 is less than half of its maximum a positive correlation results; if more than half there will be a negative correlation.

While the equation given above is not difficult to use, a version which looks rather different is more commonly used. This relies on the fact that the maximum value of Σd^2 can be worked out directly from N, the number of things ranked.

$$\text{Spearman's rho } (\rho) = 1 - \frac{6\,\Sigma d^2}{N(N^2 - 1)}$$

where Σd^2 is the sum of squared differences in rank and N = the number of pairs of ranks.

Note, by the way, that the '6' in the formula is, rather surprisingly, derived from the algebra, and is always there no matter how many pairs of scores you have.

Spearman's rho is a descriptive statistic. It simply describes and summarizes the direction and degree of the relationship between the variables. It is, however, possible to assess the statistical significance of the relationship between the variables in a similar way to that discussed with the sign test (p. 37). Table C gives figures for the smallest values of Spearman's rho significant at the 0·05 level of significance, for different numbers of pairs of scores.

If ρ exceeds the table value for the number of pairs of scores in the experiment, then there is a statistically significant agreement between the rankings under the two conditions (at the $p = 0{\cdot}05$ level). If ρ does not exceed the table value, then there is no significant agreement between the rankings under the two conditions (at the $p = 0{\cdot}05$ level).

Step-by-step procedure

Spearman's rho

Step 1 Rank data (for each group separately) giving rank 1 to the highest score, and so on
Note If two or more scores in a group are the same then give the average rank for these tied scores

Step 2 Obtain the difference (d) between each pair of ranks — d

Step 3 Square each of the differences — d^2

Step 4 Add all the squares together — Σd^2

Step 5 Calculate $N \times (N^2 - 1)$ where N is the number of pairs of scores — $N(N^2 - 1)$

Step 6 $$\rho = 1 - \frac{6\Sigma d^2}{N(N^2 - 1)}$$

Step 7 If required, assess the significance of ρ using Table C

Step 8 Translate the result back in terms of the experiment.

Note The procedure can be followed if the data is given directly in the form of ranks by simply omitting Step 1

Worked example

Spearman's rho

Participant	*Scores* *A*	*B*
P_1	3	5
P_2	7	9
P_3	3	7
P_4	12	11
P_5	8	11
P_6	14	11
P_7	2	2

Step 1 *Participant*	*ranks* *A*	*B*	**Step 2** *d*	**Step 3** d^2
P_1	5·5	6	−0·5	0·25
P_2	4	4	0	0
P_3	5·5	5	0·5	0·25
P_4	2	2	0	0
P_5	3	2	1	1
P_6	1	2	−1	1
P_7	7	7	0	0

Step 4 $\Sigma d^2 = 2{\cdot}50$

Step 5 $N \times (N^2 - 1) = 7 \times (49 - 1) = 336$

Step 6

$$\rho = 1 - \frac{6\,\Sigma\, d^2}{N(N^2 - 1)} = 1 - \frac{6 \times 2{\cdot}50}{336}$$

$$= 1 - 0{\cdot}045 = +0{\cdot}955$$

Step 7 From Table C, ρ must exceed 0·71 for $N = 7$. As $\rho = 0{\cdot}955$ there is a statistically significant agreement between the orderings of the data at the $p = 0{\cdot}05$ level

Step 8 There is a statistically significant positive correlation ($\rho = +0{\cdot}955$) between the scores on the two variables

Note Spearman's rho should be treated with caution when there is a high proportion of ties as in this example