Spearman's rank correlation

Introduction

Rank correlation is used quite extensively in school subjects other than mathematics, particularly geography and biology. There are two accepted measures of rank correlation, Spearman's and Kendall's; of these, Spearman's is the more widely used. This paper is written to give teachers of mathematics the background knowledge they need to be able to answer questions from staff and students in other departments. This can raise difficulties that do not arise in the controlled environment of the mathematics classroom. Advice is often sought only after the data have been collected, rather than when an experiment is being designed.

It will also provide useful information for teachers of the MEI S2 and Z3 modules. Examiners' reports show that candidates sometimes fail to state hypotheses for tests of correlation appropriately and they do not always connect the results of hypothesis testing to the original situation.

Bivariate data

The data referred to in this paper are all bivariate. So each data item is reported in terms of the values of two attributes. These could, for example, be the heights and weights of 11-year old girls. In keeping with common convention, the two variables are referred to separately as *X*, with sample values $x_1, x_2, \ldots, x_n$, and *Y*, with sample values $y_1, y_2, \ldots, y_n$, or together as the bivariate distribution (X, Y) with sample values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. A general bivariate item is denoted by (x_i, y_i). Sample size is denoted by *n*.

Association and correlation

The terms association and correlation are often used interchangeably; this is not correct. There is *association* between two variables if knowing the value of one provides information about the likely value of the other. There is *correlation* between the variables if the association is linear; this can be represented by a straight line on a scatter diagram. These situations are illustrated in Figures 1 and 2. People sometimes describe situations like that in Figure 2 as "non-linear correlation" but this is technically incorrect; the correct description would be "non-linear association" or just "association".

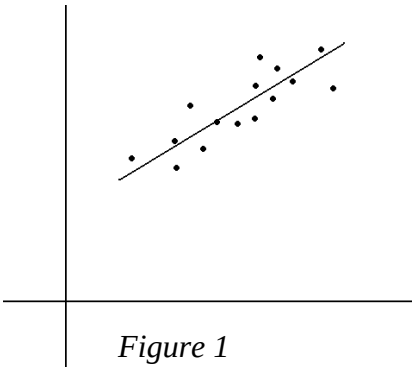

Figure 1

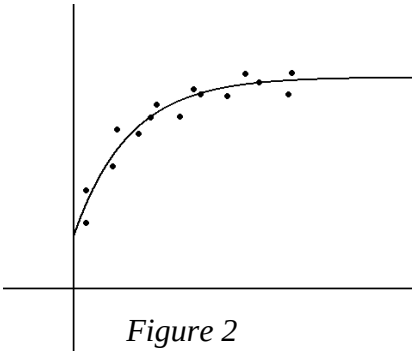

Figure 2

In the linear case, the strength of the association can be measured by the correlation coefficient; the closer the points to the straight line, the stronger is the correlation. A common mistake is to think that the steeper the line the better the correlation but this is not the case.

Product moment correlation

A commonly used measure of correlation is provided by Pearson's product moment correlation coefficient (pmcc). This is denoted by *r* and calculated from sample data using the formula

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

where $S_{xx} = \sum(x_i - \bar{x})^2$, $S_{yy} = \sum(y_i - \bar{y})^2$ and $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$.

The pmcc also provides a test statistic for the null hypothesis that there is no correlation between the two variables in the bivariate parent population from which the sample data were drawn. For the pmcc test, both the variables must be random. The notation for the population correlation coefficient is ρ, the equivalent Greek letter to *r* which is used for the test statistic.

It is usual to state the null hypothesis as

"H_0: There is no correlation, $\rho = 0$".

Since ρ is a parameter of the (bivariate) population, the inclusion of the statement "$\rho = 0$" emphasises the point that, as is standard procedure for hypothesis testing, the test is being carried out on the parent population. It is good practice to emphasise this point by including the word "population" in the statement, making it

"H_0: $\rho = 0$, where ρ is the population correlation coefficient".

Critical values for this test are provided in tables, including those in the Companion to Advanced Mathematics and Statistics.

These tables are derived from a particular standard model, the bivariate Normal distribution. So when you use the tables you are implicitly assuming that the bivariate distribution you are working with is bivariate Normal, or sufficiently close to it. A property of a bivariate Normal population is that, for any value of one variable, the distribution of the values of the other is Normal.

It is natural to ask if there is a simple way of assessing whether the distribution you are working with is bivariate Normal, or sufficiently close to it. A scatter diagram, which should usually be the first thing to be looked at anyway, provides a good guide. Among the points that a scatter diagram highlights are the following, both of which can be seen in Figure 3.

- The points representing sample data drawn from a bivariate Normal population typically form a rough ellipse.
- A bivariate Normal population may, or may not, be correlated. Where there is correlation, the axes of the ellipse are at an angle to both the *x*- and *y*-axes.

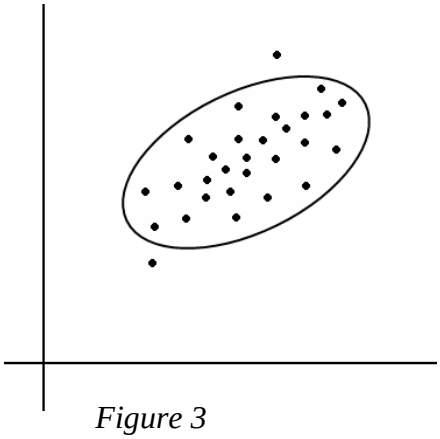

Figure 3

A sample of 100 data points from a bivariate Normal population, with $\rho = 0$, is illustrated in the scatter diagram in Figure 4. By contrast, Figure 5 illustrates a sample drawn from a bivariate Normal distribution with $\rho = 0.6$.

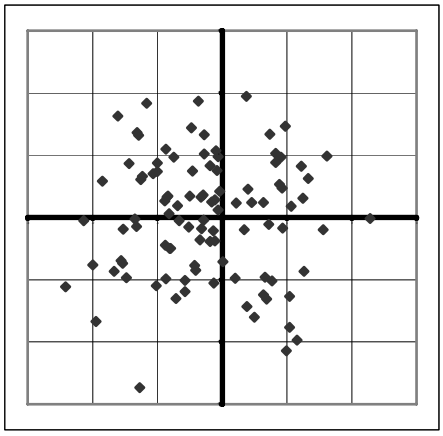

Figure 4

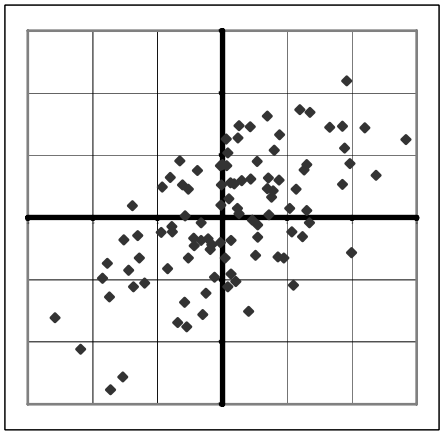

Figure 5

Rank correlation

In cases where the association is non-linear, the relationship can sometimes be transformed into a linear one by using the ranks of the items rather than their actual values.

Using ranks rather than data values produces two new variables (the ranks). These fulfil the conditions for the use of the description "correlation" since their relationship is linear but you cannot just use the pmcc test on the ranks because they are not drawn from a bivariate Normal population.

There are also situations when the only data available are in the form of ranks, for example when judges place contestants in order, 1st, 2nd, 3rd , ... but do not assign any other data values, such as scores or points, to them.

Spearman's coefficient of rank correlation, denoted by r_s, can be calculated by applying the formula for the pmcc to the ranks, although it is more usual to use the equivalent, but more algorithmic, formula

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i is the difference in the ranks given to the two variable values for each item of data. (This is only an equivalent formula when there are no tied ranks but, if there are only a few tied ranks, it provides a sufficiently good approximation.)

Because the data (i.e. the ranks) used in the Spearman test are not drawn from a bivariate Normal population, the tables of critical values are worked out differently from those for the pmcc and, hence, have different values.

The Spearman tables are constructed by considering all possible cases. Imagine that you are dealing with samples of size 10, and that you rank the data items from 1 to 10 according to the value of the *X* variable; then the corresponding ranks for the *Y* variable will be the numbers 1 to 10 in some order. One possible order is shown in Table 6.

X -rank	1	2	3	4	5	6	7	8	9	10
Y -rank	5	8	2	1	10	4	9	3	6	7

Table 6

For the particular arrangement in Table 6, the value of r_s works out to be 0.1636, but this is only 1 among 10! = 3 628 800 possible ways in which the numbers 1 to 10 can be assigned to the *Y*-rank variable. Each of these gives rise to a value of r_s, giving a distribution of possible values between -1 and +1; the critical values are worked out from this distribution. So for a 2-tail test at the 10% significance level, the critical values separate off the 5% tail nearest to -1 and the equivalent 5% tail nearest to +1. Since the distribution is symmetrical about zero, only the positive value is given in the tables.

Conditions underlying the Spearman test

The Spearman test uses ranks to test for association. However, association is a wide term covering many different types of relationship and not all of these will be picked up by the Spearman test. Figure 7 illustrates a form of association in which the points on the scatter diagram form an inverted letter U and this would not produce a significant test result.

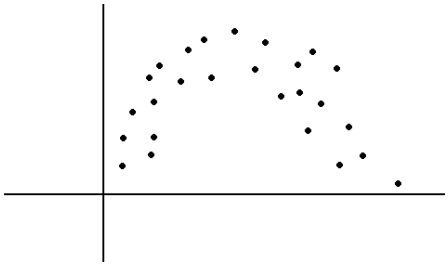

Figure 7

For the Spearman test to work, the underlying relationship must be monotonic: that is, either the variables increase in value together, or one decreases when the other increases.

Whereas using the pmcc involves the assumption that the underlying distribution is bivariate Normal, no such assumption is needed for the Spearman test. Like other procedures based on ranks, Spearman's test is non-parametric. Non-parametric tests are sometimes described as "distribution free". Spearman's test does not depend on the assumption of an underlying bivariate Normal distribution (with parameters μ_X, μ_Y, σ_X, σ_Y and ρ) or any other distribution.

Using rank correlation to test for association

As with any other hypothesis test, for Spearman's test you take a sample, work out the test statistic from the sample and compare it to the critical value appropriate for the sample size, the required significance level and whether the test is 1- or 2-tail.

The null hypothesis should be written in terms of there being no association between the variables. This conveys the purpose of the test: investigating possible association in the underlying population. It may be tempting to write the null hypothesis as "There is no rank correlation between the variables" but that would be incorrect and could be confusing. There may be a level of rank correlation for the sample but the purpose of the test is to make an inference about the population.

Difficulties arise if you try to complete the statement in symbols, the equivalent of $\rho = 0$ in the pmcc test. As has already been mentioned the term association covers many different kinds of relationship between the two variables so it is not surprising that there is no recognised notation for a measure of association that is not correlation.

In this situation some people write $\rho = 0$ but this confuses different situations. So it is better to avoid a symbolic statement of the null hypothesis altogether and instead to use words to make the point that the test is for association in the population by adding the words "in the underlying population" to the end of the statement about the null hypothesis, making it something like:

H_0: There is no association between the variables in the underlying population.

You would usually express this in a way that is appropriate to the test being carried out.

A common context for questions on Spearman's test involves two judges ranking contestants in a competition. It is important for students to realise that the test is of their underlying judgements: are they looking for different things? In such questions there is almost always a measure of disagreement over the rankings of the contestants and so it is self-evident that the judgements for the particular sample are not the same; you do not need a test to tell you that. It is the underlying nature of the judgements that is being examined. In this context, the null hypothesis would be:

H_0: There is no association between the judgements of the two referees.

Sample size

There are difficulties associated with using Spearman's test with data from either very small samples or large samples.

Very small samples

Take the case of a sample of size 4. The critical values are worked out on the 4! = 24 possible ways of arranging the *Y*-ranks. Some of these give rise to the same values of d^2 and so there are fewer than 24 possible values of r_s for samples of this size. The actual distribution is shown below in Table 7. (Figures are given to 3 decimal places.)

r_s	-1	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1.0
p	0.042	0.125	0.042	0.167	0.083	0.083	0.083	0.167	0.042	0.125	0.042

Table 7

Table 7 illustrates two important points.

- The distribution of r_s is discrete. In this case, with a sample of size 4, there are intervals of 0.2 between the possible values of r_s. Smaller samples have larger intervals between the possible values, this limits the precision which is possible in setting critical values, particularly for small samples.

- The probabilities of values of 1 and –1 are 4.2%. These would be caught by significance levels of 5% (1-tail) or 10% (2-tail). More sensitive tests are not possible with samples of size 4.

These points are examples of the difficulties of using Spearman's test with very small samples. In such cases, the test should be used with caution, taking care not to over-interpret the outcomes.

Large samples

Different problems arise when you try to use Spearman's rank correlation on larger samples.

The first point is that it becomes very time-consuming because of the need to rank the data for both variables.

The tables of critical values are given only for a limited range of possible sample sizes; those in the Companion to Advanced Mathematics and Statistics go up to $n = 60$. This problem is not insuperable because of two possible approximations for larger samples:

- For values of *n* from about 20 upwards, $r_s\sqrt{\frac{n-2}{1-r_s^2}}$ has an approximate *t* distribution with $n-2$ degrees of freedom
- For values of *n* from about 40 upwards, $r_s\sqrt{n-1}$ is approximately N(0,1).

Which is the better test for correlation ?

There are data sets which could be tested for correlation using either the pmcc test or Spearman's test. Which is better? To answer this question, look at the data in Table 8, which shows the heights and weights of twenty children.

Height (cm)	89	88.8	90.9	87.4	88.7	90.8	92.5	92.2	88.4	94
Weight (kg)	11.7	11.8	12	12.2	12.4	12.5	12.6	12.7	12.8	13
Height rank	5	4	11	1	3	10	15	13	2	18
Weight rank	1	2	3	4	5	6	7	8	9	10

Height (cm)	92.1	90.5	89.8	94.1	89.2	92.4	95.1	90.7	93.9	93.8
Weight (kg)	13.1	13.2	13.3	13.4	13.5	13.6	13.7	13.8	14	14.4
Height rank	12	8	7	19	6	14	20	9	17	16
Weight rank	11	12	13	14	15	16	17	18	19	20

Table 8

The data are plotted on a scatter diagram in Figure 9, on the next page.

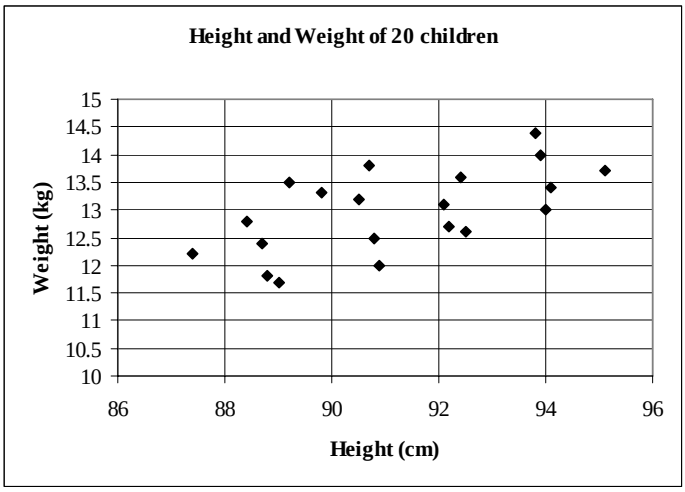

Figure 9

It looks as though the sample could have been drawn from a bivariate Normal population, so you could use either test for correlation.

When testing for correlation at the 1% level of significance (2-tail test), the two tests give different outcomes:

- Pearson's pmcc is 0.611 (3dp), the critical value is 0.5614 so the alternative hypothesis, that there is correlation, would be accepted.
- Spearman's rank correlation coefficient is 0.558 (3dp), the critical value is 0.5699 so the null hypothesis, that there is no association in the underlying bivariate population, would be accepted.

At first sight this is very confusing. You must however remember two things that make it less worrying that statistical procedures can give different results. First, the null hypotheses being tested are different: in the Pearson case the null hypothesis is strictly that that there is no correlation in the underlying bivariate population, whereas Spearman's test is looking at the rather more diffuse null hypothesis of no association. Secondly, there is the question of the sensitivities of the tests and how strongly they rely on background assumptions. The Pearson test may be expected to be more sensitive (this is referred to as more *powerful* in more advanced statistical work) if the assumption of bivariate Normality is correct, but the Spearman test gives rather more insurance in case this assumption is not correct.