Spearman’s Rank Correlation Coefficient.

Tests of Correlation

When we are considering the question of a possible correlation between two observed physical quantities we require:

(1) a precise mathematical determination of the degree of correlation;
(2) a statement of the probability that the correlation could arise by chance.

Ordinal data[1] are measures of physical quantities that can be ranked. For example, the variable X could measure the number of days individuals have been subject to a special diet; the variable Y could measure the position of those individuals in a race. Here, it is meaningful to ask how does the position of an individual, that is his rank, in terms of values of X correlate with his position, or rank, in terms of Y?

Spearman’s Coefficient of Rank Correlation

Let X and Y be two variables at ordinal data level. Let rank X represent the order in which the values of X occur, and likewise rank Y represent the corresponding order in which values of Y occur. Each value of X is associated with a value of Y – they form pairs of values. Then, let

$$d = \text{rank}\,X - \text{rank}Y$$

Then, Spearman’s rank correlation coefficient is:

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

where n is the number of pairs of values $\left(X_i, Y_i\right)$.

Example

[1] For further discussion of the meaning of ordinal data and comparison with interval data see articles on *Tests of Correlation* and *Research Statistics: Choosing a Test.*

Two independent examiners awarded the following marks to each of five candidates:

Candidate	A	B	C	D	E
1st Examiner	39	90	64	81	60
2nd Examiner	53	84	42	85	61

Calculate Spearman's rank correlation coefficient for these data.

Note: The data is at ordinal level because it is not meaningful to say that the difference between a score of 85 and a score of 80 is equivalent, for example, to the difference between 45 and 40.

Candidate	1st	2nd	Rank	Rank	d	d^2
A	39	53	5	4	1	1
B	90	84	1	2	-1	1
C	64	42	3	5	-2	4
D	81	85	2	1	1	1
E	60	61	4	3	1	1
						$\sum d^2 = 8$

Therefore,

$$\begin{aligned} r_s &= 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} \\ &= 1 - \frac{6 \times 8}{5 \times 24} \\ &= 0.6 \end{aligned}$$

Hypothesis testing and Spearman's rank correlation coefficient

The value of the correlation coefficient *suggests* whether there is a linear relationship or not. A high value of r indicates a strong positive relationship. However, there is the question of just how *likely* a particular value for the product moment correlation coefficient could be.

The apparent correlation between two variables X and Y could be due to chance factors alone. The question must be, what is the probability that X and Y could appear to be correlated with a coefficient of correlation r due to chance factors alone?

This leads to hypothesis testing of the correlation coefficient.

We seek to test the hypothesis that there is a correlation, that is $r \neq 0$ against the null hypothesis that there is no correlation, $r = 0$.

Such a test would be formulated as

$\mathrm{H}_0 \quad r = 0$

$\mathrm{H}_1 \quad r \neq 0$

This is called a *two-tailed test* because the correlation could be either positive, $r > 0$, or negative, $r < 0$. On the other hand, we might believe in advance of collecting the data that there was either a positive or a negative correlation, in which case we would be testing, in the case of a positive correlation

$\mathrm{H}_0 \quad r = 0$

$\mathrm{H}_1 \quad r > 0$

or in the case of a negative correlation

$\mathrm{H}_0 \quad r = 0$

$\mathrm{H}_1 \quad r < 0$

In both these cases, the test is a *one-tailed test* because we are expecting either a positive correlation *or* a negative correlation, but not both.

When we are testing a hypothesis using Pearson's product moment correlation coefficient[2] we have to assume that the two variables are jointly normally distributed.

[2] For explanations of these terms see articles on *Tests of Correlation* and the *Standardised Normal Variable*.

However, with Spearman’s rank correlation coefficient, no such assumption has to be made.

Hypothesis tests are made at a *significance* level. This indicates the degree of chance that you are willing to accept and yet allow the data to pass the hypothesis test. For example, if the significance level is 5% then you are allowing that on 5% of the occasions the data would appear to be correlated *by chance* when in reality it is not correlated. This corresponds to a probability of 1 in 20. In other words, you are accepting that there is a correlation, even though there is a 1 in 20 chance that the same degree of correlation would appear merely due to chance alone.

In order to perform the test you require a set of tables of *critical values* for the Spearman’s rank correlation coefficient. These tables give you the critical values of the correlation coefficient for a given number of data and a given significance level. We will illustrate this by means of a complete worked example.

Example

(a) A child is asked to place 12 objects in order and gives the ordering

D A B G K E C I L F J H.

The correct ordering is

A B C D E F G H I J K L.

Find a coefficient of rank correlation between the child’s ordering and the correct ordering.

(b) Test at the 5% significance level whether the child’s ordering is positively correlated with the correct order.

Solution

(a) The number of items is $n = 12$.

Child's ranking, x	4	1	2	7	11	5	3	9	12	6	10	8
Correct ranking, y	1	2	3	4	5	6	7	8	9	10	11	12
d	3	-1	-1	3	6	-1	-4	1	3	-4	-1	-4
d^2	9	1	1	9	36	1	16	1	9	16	1	16

$$\sum d^2 = 116$$

The test statistics is given by

$$r_{test} = 1 - \frac{6\sum d^2}{n\left(n^2 - 1\right)} = 1 - \frac{6 \times 116}{12 \times \left(144 - 1\right)} = 0.5944$$

(b) We are testing for whether there is a positive correlation. The hypothesis, H_1 is that there is a positive correlation; the null hypothesis, H_0 is that there is no correlation. In symbols

$$H_0 : \quad r = 0$$
$$H_1 : \quad r > 0.$$

The significance level is $\alpha = 0.05$, and this is a one tailed-test.

From tables the critical value for a significance level of 0.05 and n = 12 is $r_{\text{critical}} = 0.5035$. ($\alpha = 0.05,\ n = 12$, one-tailed test.)

The value of the test statistic is $r_{test} = 0.5944 < 0.5035 = r_{critical}$

Therefore, accept H_1 and reject H_0 at 5% level of significance. There is evidence that the child's ordering is positively correlated with the correct one.

(You accept the hypothesis if the test value is greater than the critical value.)

Derivation of the Spearman rank correlation coefficient

This derivation depends on your being familiar with Pearson's product moment correlation coefficient and the concept of covariance.[3] Theoretically the rank correlation coefficient is derived in exactly the same way as the Pearson's product moment correlation coefficient, only the data are ranks and not values.

[3] For this see *Pearson's Product Moment Correlation Coefficient*

Thus

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

where r is either Spearman's rank correlation coefficient or Pearson's product moment correlation coefficient.

However, when ranks are used a considerable simplification of the formula for calculating the coefficient is possible. As we have seen it is simply given by

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

We shall derive this formula from

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

To prove this, firstly, let us assume that there no tied ranks, meaning no ranks with the same score.

Suppose there are n pairs of values. Then $x_1, x_2 ... x_n$ are the numbers 1,2,…n in some order and likewise $y_1, y_2 ... y_n$ are the numbers 1,2,…n in some order.

Then $\sum x = \sum y = 1 + 2 + ... + n$ which is an arithmetic series with n terms and common difference 1.

Hence $\sum x = \sum y = \frac{n}{2}\left(2 \times 1 + \left(n-1\right) \times 1\right) = \frac{n}{2}\left(n+1\right)$

Also $\sum x^2 = \sum y^2 = \frac{n}{6}\left(n+1\right)\left(2n+1\right)$

We will assume this result, which is proven by mathematical induction, and is a standard result for a series.[4]

Hence

$$
\begin{aligned}
S_{xx} &= \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \\
&= \frac{n}{6}(n+1)(2n+1) - \frac{\left(\frac{n}{2}(n+1)\right)^2}{n} \\
&= \frac{n}{6}(n+1)(2n+1) - \frac{n(n+1)^2}{4} \\
&= \frac{2n\left(2n^2+3n+1\right) - 3n\left(n^2+2n+1\right)}{12} \\
&= \frac{n\left(4n^2+6n+2-3n^2-6n-3\right)}{12} \\
&= \frac{n\left(n^2-1\right)}{12}
\end{aligned}
$$

Likewise $S_{yy} = \sum y_i^{\;2} - \frac{\left(\sum y_i\right)^2}{n} = n\frac{\left(n^2-1\right)}{12}$

We need to seek a formula for S_{xy}. To find this, note that

$$
\begin{aligned}
S_{xy} &= \sum (x_i - y_i)^2 \\
&= \sum \left(x_i^2 + y_i^2 - 2x_i y_i\right) \\
&= \sum x_i^2 + \sum y_i^2 - \sum 2x_i y_i
\end{aligned}
$$

That is

$$
\sum (x_i - y_i)^2 = \sum x_i^2 + \sum y_i^2 - \sum 2x_i y_i
$$

Rearrangement of this formula gives

[4] For this result see *Series by Standard Results* and *Mathematical Induction*

$$\sum 2x_i y_i = \sum x_i^2 + \sum y_i^2 - \sum \left(x_i - y_i\right)^2$$

$$\sum x_i y_i = \frac{\sum x_i^2 + \sum y_i^2 - \sum \left(x_i - y_i\right)^2}{2} \tag{1}$$

Now we have already seen that

$$\sum x_i^2 = \frac{n}{6}\left(n+1\right)\left(2n+1\right), \qquad \sum y_i^2 = \frac{n}{6}\left(n+1\right)\left(2n+1\right)$$

Hence

$$\frac{\sum x_i^2 + \sum y_i^2}{2} = \frac{n}{6}\left(n+1\right)\left(2n+1\right)$$

And on substituting into equation $\left(1\right)$ above

$$\sum x_i y_i = \frac{n}{6}\left(n+1\right)\left(2n+1\right) - \frac{1}{2}\sum \left(x_i - y_i\right)^2$$

Hence, given the standard formula for calculating S_{xy}

$$\begin{aligned}
S_{xy} &= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \\
&= \frac{n}{6}\left(n+1\right)\left(2n+1\right) - \frac{1}{2}\sum \left(x_1 - y_i\right)^2 - \left(\frac{n}{2}\left(n+1\right)\right)\left(\frac{n}{2}\left(n+1\right)\right) \\
&= \frac{n}{6}\left(n+1\right)\left(2n+1\right) - \frac{1}{2}\sum \left(x_1 - y_i\right)^2 - \frac{n^2\left(n+1\right)^2}{4n} \\
&= \frac{n\left(n^2-1\right)}{12} - \frac{1}{2}\sum \left(x_i - y_i\right)^2
\end{aligned}$$

Hence

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

$$= \frac{\dfrac{n\left(n^2-1\right)}{12} - \dfrac{1}{2}\sum\left(x_i - y_i\right)^2}{\sqrt{\left(\dfrac{n\left(n^2-1\right)}{12}\right) \times \left(\dfrac{n\left(n^2-1\right)}{12}\right)}}$$

$$= \frac{\dfrac{n}{12}\left(n^2-1\right) - \dfrac{1}{2}\sum\left(x_i - y_i\right)^2}{\dfrac{n}{12}\left(n^2-1\right)}$$

$$= 1 - \frac{6\sum\left(x_c - y_i\right)^2}{n\left(n^2-1\right)}$$

The term $x_i - y_i$ is the difference between the two pairs of values. letting $d_i = x_i - y_i$ then

$$r = 1 - \frac{6\sum d^2}{n\left(n^2-1\right)}$$

This is the form taken by Spearman’s rank correlation coefficient.