Bio Factsheet

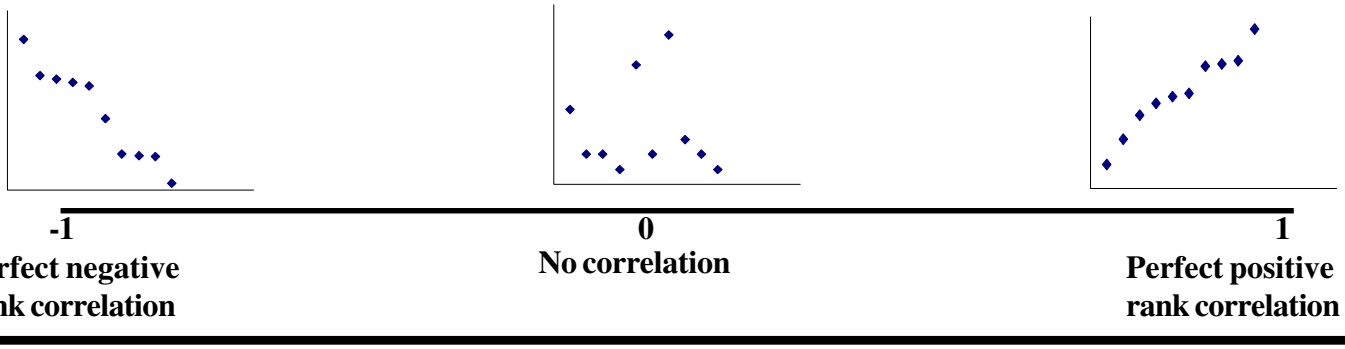
www.curriculum-press.co.uk Number 144

Spearman's Rank Correlation Coefficient

Spearman's Rank is one method of measuring the correlation between two variables. Correlation may be:

- **positive (large values of one variable associated with large values of the other variable - eg nitrate concentration and plant growth)**
- **negative (large values of one variable associated with small values of the other - eg soil salinity and plant growth)**

Correlation is measured on a scale from -1 to 1



Which correlation coefficient?

There are three correlation coefficients in common use; Spearman's is used most often (and hence is the principal subject of this Factsheet), but there are cases when the other coefficients should be conisdered:

Spearman's Rank Correlation Coefficient

- Can be used for any data that you can put in order smallest to largest
- Measures whether data are in the same **order** - eg does highest nitrate concentration coincide with highest plant growth - rather than using actual data values
- Not valid if there are a lot of ties (eg several pairs of samples having the same pollution level), although one or two ties is OK.
- Easy to calculate for small data sets, but unwieldy for large data sets.

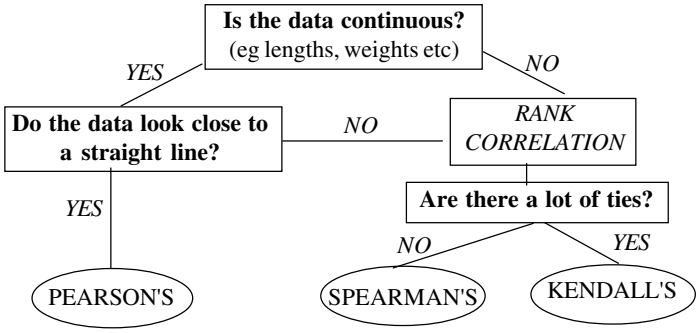
Kendall's Rank Correlation Coefficient

- Like Spearman's, uses the ranks of the data rather than the actual data, and can be used for any data that can be ordered.
- A good substitute for Spearman's if there are a lot of ties
- More of a nuisance to calculate than Spearman's

Pearson's Product Moment Correlation Coefficient

- Can only be used for continuous data (eg lengths, weights)
- Uses the actual data, not just their ranks
- Measures how close to a straight line the data are - check on a scatter graph that the data do approximate a straight line rather than a curve.
- Can be easier to get significant results than using rank correlation
- A nuisance to calculate by hand, but can be calculated automatically on many graphic calculators and using a spreadsheet
- If you are unsure whether it is valid, it's better to use rank correlation

The flowchart shows how to choose your correlation coefficient.



Hypotheses

As with any other statistical test, you are using the test to decide between two hypotheses: -

- the **null hypothesis (H_0)** - which is what you assume, until you get convincing evidence otherwise.
- the **alternative hypothesis (H_1)** - which is what you hope to get evidence for.

For any test of correlation, your null hypothesis is **always**:

H_0: there is no correlation between X and Y

The alternative hypothesis can take three possible forms:

a) H_1: there is some correlation between X and Y or

b) H_1: there is positive correlation between X and Y or

c) H_1: there is negative correlation between X and Y

If you have a good scientific reason in advance (before actually getting any results) for expecting a particular type of correlation, then choose b) or c). If you do not have a reason for expecting a particular type, use a). *If in doubt - use a)*

Alternative hypotheses b) and c) above are referrred to as **directional** - because they specify a particular "direction" of correlation. Alternative a) is **non-directional.** When you are doing the actual statistical test, you need to be aware that a non-directional alternative requires you to do a **2-tailed test**, but a directional alternative requires a **1-tailed test** - further details are given in the worked example overleaf.

Exam Hint: - *Only the **alternative** hypothesis can be directional - the null hypothesis is never directional.*

Sample Size

The absolute minimum number of values for using Spearman's Rank is 4 - but it is very hard to get a significant result using this few! It's best to use at least 7 - and if you can get up to about 15, better still. Very large sample sizes (50+) can make it hard to handle the calculations, and many Spearman's tables do not go up this high.

Ranking

Ranking is similar (though not identical) to awarding places in a race. When doing the ranking, it does not matter whether you give the rank "1" to the largest value, or to the smallest value - **provided you are consistent.**

If there are no ties, you just give out the ranks in the obvious way, starting at 1 and carrying on to however many pieces of data you have.

If there are ties, you have to be a bit careful:

For example, suppose three pieces of data tie for 4th place.

Normally, if there hadn't been any ties, you'd expect the next three pieces of data to "use up" the ranks 4, 5, 6

So we give all three pieces the average of 4, 5 and 6 - that's 5.

The next piece of data then has rank 7 (as ranks 4, 5 and 6 have been "used up")

Worked Example

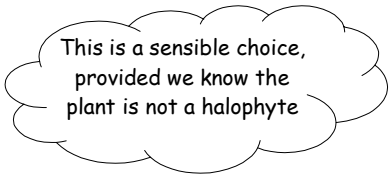
The data below were collected on soil salinity and plant height.

Soil Salinity	28	12	15	16	2	5
Plant height (mm)	10	40	40	52	75	48

Step 1: Write down the **hypotheses**

H_0: There is no correlation between soil salinity and plant height

H_1: There is negative correlation between soil salinity and plant height

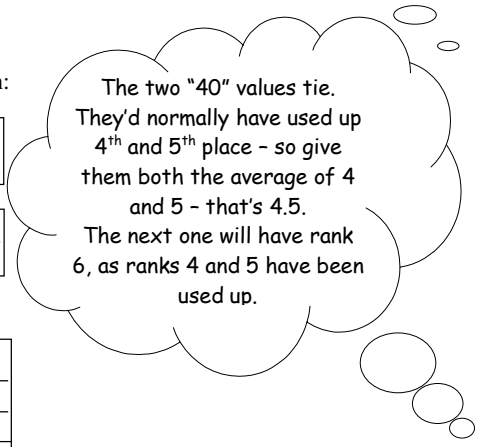


Step 2: Work out the two sets of **ranks**, taking care to allow for ties.

We'll give rank 1 to the highest values for each:

Soil Salinity	28	12	15	16	2	5
Rank	***1***	***4***	***3***	***2***	***6***	***5***

Plant height (mm)	10	40	40	52	75	48
Rank	***6***	***4.5***	***4.5***	***2***	***1***	***3***



Step 3: Work out "d" and "d^2", where d stands for the differences between pairs of ranks

Note: you must square each d individually

Soil salinity rank	1	4	3	2	6	5
Plant height rank	6	4.5	4.5	2	1	3
d	***5***	***0.5***	***1.5***	***0***	***5***	***2***
d^2	***25***	***0.25***	***2.25***	***0***	***25***	***4***

Step 4: Substitute into the formula

$$r_s = 1 - \frac{6\Sigma d^2}{(n^3 - n)}$$

$\Sigma d^2 = 25 + 0.25 + 2.25 + 0 + 25 + 4 = 56.5$ $\quad n = 6$

$$r_s = 1 - \frac{6 \times 56.5}{(6^3 - 6)} = 1 - \frac{339}{210} = -0.6142$$

r_s = Spearman's Rank Correlation Coefficient
Σd^2 = sum of the d^2 values
n = number of pairs of values in sample

Step 5: Get a Spearman's table and **look up the critical value** for the appropriate significance level (usually 5% = 0.05), sample size and 1-tailed or 2-tailed test.

We have n = 6, and we are doing a one-tailed test, because of the form of H_1.

So critical value is **0.771**

1-tail	0.1	**0.05**	0.025	0.01	0.005
2-tail	0.2	0.10	0.05	0.02	0.01
n					
4	1.000	1.000	1.000	1.000	1.000
5	0.700	0.900	0.900	1.000	1.000
6	0.657	0.771	***0.829***	0.943	0.943
7	0.571	0.679	0.786	0.857	0.893

Step 6: **Make a decision** - if your calculated chi-squared value is **bigger** than the critical value (ignoring signs), you can **reject** the null hypothesis. Otherwise you must accept it.

Our value (-0.6142) is **smaller** than the critical value (ignoring signs)
So we must **accept** the null hypothesis - there is no correlation between soil salinity and plant growth at the 5% significance level.

Further Investigations Using This Test

- Relationship between concentration of fungicide and zone of inhibition for a particular fungus
- Relationship between molecular size and rate of metabolism in yeast
- Relationship between algal growth and nitrate concentration
- Relationship between blackspot disease in roses and traffic levels
- Relationship between mass of leaf buried and earthworm mass
- Relationship between pest density and yield for broad beans
- Relationship between body mass and running ability for house spider
- Relationship between pH of soil and pH of leaf litter

Acknowledgements: *This Bio Factsheet was researched and written by Cath Brown.* ***Curriculum Press.*** *Bank House, 105 King Street, Wellington, TF1 1NU.* ***ISSN 1351-5136***