Geofile Online
JANUARY 2006
511

ALISON RAE

USING SPEARMAN'S RANK CORRELATION COEFFICIENT IN COURSEWORK

Introduction – use of statistical methods in coursework

Statistical techniques can often be used in AS and A2 level coursework to add a more analytical dimension and hence raise the mark. Some students are so keen to use these techniques that they overdo things and may use them inappropriately. Other students tend to be under-confident in the use of statistical techniques, especially those who are not studying mathematics. A balance therefore needs to be struck, being aware of the following:

- that the technique/s chosen is/are relevant to the study
- that too many techniques are not used
- that the technique/s is/are used accurately
- that the results of the statistical process/es are carefully integrated into the analysis section of the study's write-up.

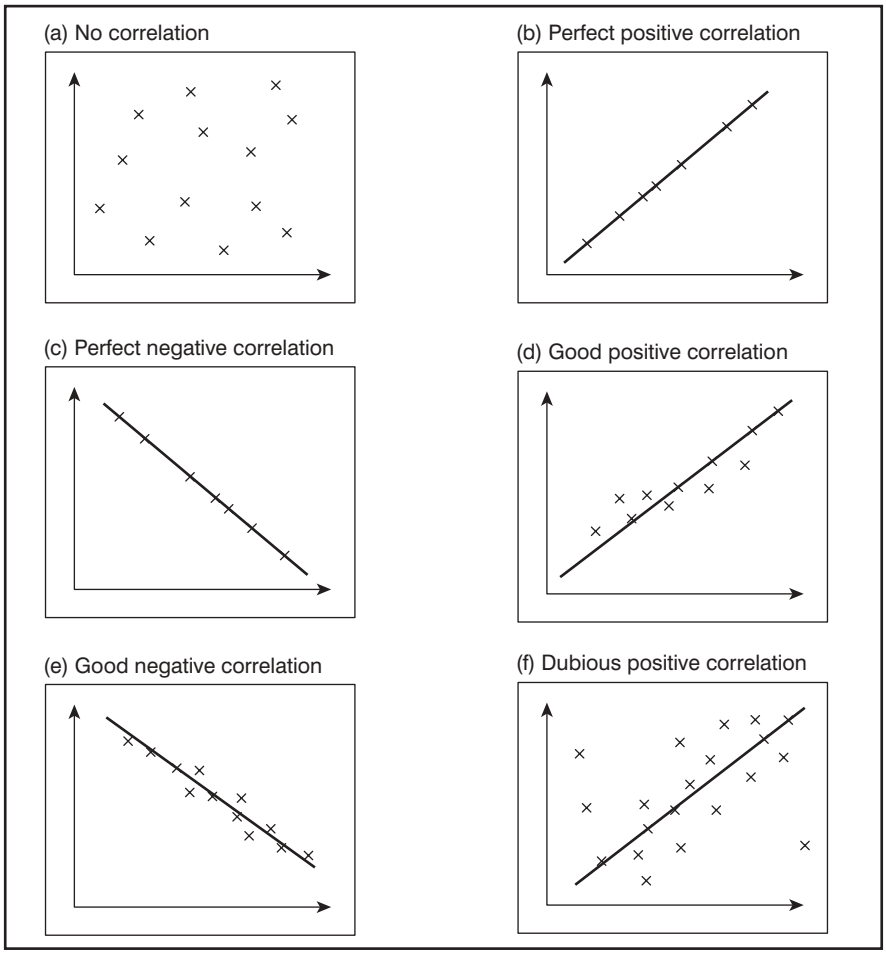
Remember, it is better not to use statistics at all than to misuse them.

The relevance of the Spearman method

A level coursework, particularly that for AS, often involves statistical techniques which reveal the degree of similarity or difference between two data sets. Difference can be measured in various ways – comparing two means, or the Mann-Whitney U test (which compares the medians of two sets of figures), or the Chi-squared test.

The Spearman's Rank Correlation Coefficient method is used to assess the degree of similarity or correlation between two data sets. It is probably one of the more simple methods to use. For those students who feel rather afraid of applying statistics to their geographical data, this method is really quite straightforward. A good place to start is in understanding the basic concept of correlation.

Figure 1: Types of correlation



Correlation

Correlation can also be represented graphically by scattergrams and the drawing (or not) of a best fit line (Figure 1). Figure 1(a) shows the situation where there is no correlation at all. The points appear random, i.e. there is no trend or pattern at all and it would be completely impossible to draw any kind of best fit line. The opposite extreme of this situation is represented in Figures 1(b) and 1(c), both of which show perfect correlation, in other words, all the points are on the best fit line. This is not a situation you are very likely to encounter using data from a geographical investigation. Figures 1(d), 1(e) and 1(f) are the type of relationship you are more likely to have with your own data: 1(d) and 1(e) show a situation of strong correlation, where some of the points are on the line, though most are not but are reasonably close to it. The situation in Figure 1(f) is more of a problem. It is possible to draw a best fit line, but several points are still quite far from it. There is therefore probably some relationship between the two sets of data, but from the scattergraph it is difficult to gauge exactly how much. This is where Spearman's Rank Correlation Coefficient is valuable. It places a **numerical** value on the degree of correlation of the two lists of measurements.

Box 1 Positive and negative correlation

Positive correlation = as one set of values increases, so does the other. The best fit line on the graph is drawn from the bottom left to the top right.
Negative correlation = as one set of values increases, the other decreases.
The best fit line on the graph is drawn from the top left to the bottom right.

Box 2 Purpose of Spearman's R

The method used by Spearman to compare a paired set of ranked values in order to look for a correlation. The higher the correlation co-efficient, the greater is the correlation between the two sets of numbers.

The variables can be compared in a precise **and quantitative way**.

Philip's Geography Dictionary (1995) pp. 44, 206

Uses of Spearman's R

This statistical method can be applied where there are two sets of data relating to various points of data collection in the field. The investigator should have reason to believe that there is some sort of relationship between the data sets. This belief could come from observation of the data, i.e. the figures simply look as if there is a relationship between them, or it could be an assumption from geographical knowledge, e.g. that it is likely there will be a relationship between the distance from the source of a stream and its width at any given point.

Operating Spearman's Rank Correlation Coefficient

This process is based on the **ranks** of the individual values of two variables, and not on the actual values themselves. There must be a minimum of 10 pairs of values in the data sets – the more, the better. Below 10 sets of values, the statistical result is unreliable and you might have marks deducted in coursework for attempting the procedure in this circumstance.

The best way to approach the procedure is by the use of a table (Figure 2). Fill in your measured data in the first and third columns. Rank them in the second and fourth columns, counting the largest value as rank number 1, and so on down your list.

Where two or more of the values are of equal rank, average the ranks. If this happens with two values, omit the next ranking, e.g. rank order is 1, 2, 3, 4.5, 4.5, 6, etc.

Obviously, there should be as many ranks as there are values in the list of data. The fifth column, d, represents the difference between the ranks for each pair of values along a row in the table. After that, you need to calculate the sixth column values, d squared, simply by squaring the values in column five.

Figure 2: Table for Spearman's R calculation

1st variable	rank	2nd variable	rank	d	d^2

The table is now complete; you have all the numbers required to substitute into the Spearman's Rank Correlation Coefficient equation. The equation is:

$$R = 1 - \frac{6 \times d^2}{n(n^2 - n)}$$

where: R is the Spearman's coefficient
d^2 is the sum of all the squared differences, as calculated in the table
n is the number of pairs of data in the list in the table.

An example of the procedure is worked through in the example below.

Interpreting the result

The result will lie within the range +1 to –1. Should it not do so, a mistake has been made and the whole procedure should be reviewed to find the error. The exact place the result lies between these values is of great significance. The key values are set out below and in Figure 3:

+1	=	perfect positive correlation
–1	=	perfect negative correlation
+1 to +0.7	=	significant positive correlation
–1 to –0.7	=	significant negative correlation

Box 3 Examples of studies suitable for the application of Spearman's R

Spearman can be used in a wide variety of both physical and human fieldwork topics.

Physical topics

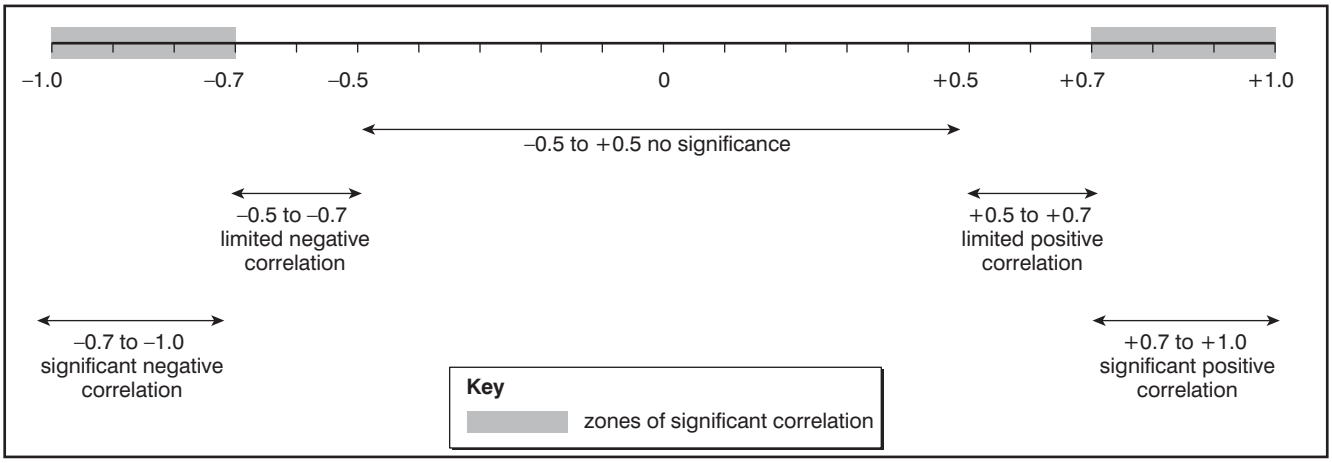
- Width of stream/river will increase with distance from source.
- Speed of river flow will increase/decrease with distance from source.
- Discharge of a river will increase with distance from source.
- Hydraulic radius of a river will increase with distance from source.
- Weight of water in soil (as a %) will increase downslope.
- Thickness of soil will increase downslope.
- % of ground covered with vegetation will increase as pH rises.
- % of ground covered with vegetation will increase with distance from the base of a single large conifer.
- % of ground covered with vegetation will increase as does no. of hours of sunshine received within an area of woodland.
- Temperature range increases with distance from the sea.
- Temperature (daily maximum) increases with distance inland.
- Temperature decreases with height above sea level.
- Mean vegetation height decreases with altitude.

Human topics

- Pedestrian activity decreases with distance from the PLVI in a CBD.
- No. of cars parked per 100 m decreases with distance from the PLVI.
- Building height decreases with distance from the PLVI/CBD.
- No. of storeys (in buildings) decreases with distance from the PLVI/CBD.
- Birth rate decreases as GNP/capita increases.
- Death rate decreases as GNP/capita increases.
- Natural increase decreases as personal income increases.
- No. of cars per household increases as population density decreases.

(All of these topics suggested as suitable for the application of Spearman's R represent good ideas for possible fieldwork/coursework.)

Figure 3: The range of significant results



+0.7 to +0.5 = weaker positive correlation
–0.7 to –0.5 = weaker negative correlation
+0.5 to –0.5 = no significant relationship between the two data sets – if this occurs it would be sensible to review the variables and the ways in which the data was collected, in order to look for error.

Basically, you are looking for a result lying between +1 and +0.7 or between –1 and –0.7. If your answer lies outside of these limits there is no valid correlation between your data sets. Should this be the case, it may be the time to review choice of methods overall.

Significance testing

The statistical reliability of the result must be checked, as there is always a possibility that chance has had an influence upon it. There are two levels at which this element of chance is usually checked, 5% and 1%. The larger the number of pairs in the data sets, the less likelihood there is of such inaccuracy. Reliability can be checked using a set of critical values tables, as shown in Figure 4.

Worked example of the Spearman's Rank Correlation Coefficient

The study undertaken involved measuring and explaining river channel characteristics in the Yorkshire Dales National Park for an AS level piece of coursework with a limit of 1,000 words. This is quite short, so any method which adds to the clarity and focuses the analysis section, such as use of relevant statistics, is very useful. Hypotheses set up were:

- the cross-sectional area increases with distance from the source;
- the velocity increases with distance from the source;
- the discharge increases with distance from the source;
- the hydraulic radius increases with distance from the source;
- the bedload size decreases with distance from the source.

Ten sites were used for data collection, this being the minimum required for statistical viability. Initially, these were identified using an OS map, and were as evenly spaced as possible. However, access was not always possible at all selected points, so the evenness of spread of the locations was not perfect. However, because the Spearman's R method of analysis uses a ranking system, the limitation of this uneven spacing was partly eradicated. The first data set comprised the distance (in km) of each site from the stream source, as measured on the OS map. In this case, the shortest distance, Site 1, was ranked as no. 1. Ranking could have been reversed without affecting the outcome, but this method seems more logical.

At each site a sample of 10 pebbles was chosen, one pebble from each of the points where depth measurements were also being taken for other hypotheses in the study. Their long axes were measured with callipers. The average pebble size at each site was calculated (Figure 5), and then the averages used as the second data set in the Spearman R calculation (Figure 6).

$$R = 1 - \frac{6 \quad d^2}{n(n^2-1)}$$

Figure 4: Critical values for Spearman's R at the 5% significance level:

n	Critical value	N	Critical Value
9	0.7000	20	0.4466
10	0.6485	21	0.4364
11	0.6162	22	0.4252
12	0.5874	23	0.4160
13	0.5604	24	0.4070
14	0.5385	25	0.3977
15	0.5214	26	0.3901
16	0.5029	27	0.3828
17	0.4877	28	0.3755
18	0.4716	29	0.3685
19	0.4596	30	0.3624

$$R = 1 - \frac{6x16.5}{10(100-1)}$$

$$= 1 - \frac{99}{(1000-10)}$$

$$= 1 - \frac{99}{990}$$

$$= 1 - 0.1$$

$$= 0.9$$

Interpretation of this result

Refer back in this **Geofile** to the earlier section on interpreting results, including Figure 3. The coefficient, 0.9, lies well within the zone of strong positive correlation. This therefore supports the hypothesis posed that 'bedload size decreases with distance from the source.' Figure 7 is a scattergraph showing the same data. A best fit line has been drawn and several points are on the line. No point is very far from the line.

A full written analysis of this hypothesis would include an attempt to explain the sites where bedload size did not fit closely with the

Figure 5: Pebble sizes for the first four sites in the river study

Pebble number	Site 1 Size in cm	Site 2 Size in cm	Site 3 Size in cm	Site 4 Size in cm
1	11.5	16	22	5.5
2	10	22	7	3
3	50	21	12	11.5
4	68	3	19	8
5	9	6	12	6
6	18	10	20	6.5
7	5	4	13.5	12.5
8	13	27	14	9
9	15	19.5	19	10
10	14.5	10	10.5	7.5
Average bedload size	21.40	13.85	14.90	7.95
Etc. for all sites.				

Figure 6: Spearman R calculation:

1st variable Distance from source (in km)	Rank	2nd variable Av size of bedload (long axis) (in cm)	Rank	D	d^2
0.1	1	21.40	1	0	0
0.51	2	13.85	3	1	1
0.98	3	14.90	2	1	1
1.5	4	7.95	4.5	0.5	0.25
2.1	5	7.95	4.5	0.5	0.25
2.6	6	4.70	7	1	1
3.0	7	3.50	9	2	4
3.5	8	3.65	8	0	0
3.95	9	5.90	6	3	9
4.55	10	2.70	10	0	0
			d^2	=	16.5

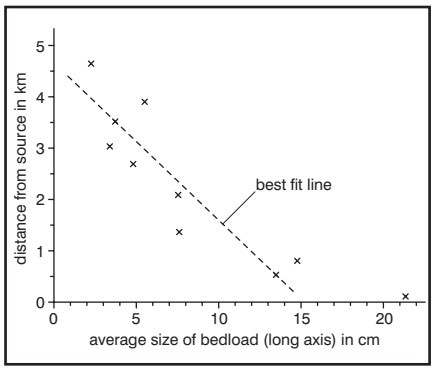
overall pattern, for instance at Site 2. Here, the valley narrows, so the same volume of water is being squeezed through a more constricted space. This leads to greater turbulence and therefore perhaps to greater erosive forces on the bedload at that point. By Site 3, the channel is a little wider again and flow is a mixture of turbulent and laminar. Also, more larger pieces of load may have fallen into the stream from the steep semi-bare rock valley sides at this location. Further investigation into the angularity/smoothness of the load would add to completeness of the study here.

This style of discussion shows how the Spearman's R statistic can be used in a geographical investigation. It is a tool to help interpret data, but it is not a replacement for wide-ranging written discussion. It can be a part of it, however.

Conclusion – advantages and limitations of Spearman's R

The main advantage of the Spearman method of assessing the degree of relationship between two data sets is that it is relatively quick. Care is needed in the interpretation of the result. The coefficient itself simply shows whether a correlation exists and, if so, how strong it is. It does not give reasons for the link. The investigator has to think broadly, considering all the evidence gathered for the study as well as geographical theory, in order to offer a full explanation of the pattern identified by the statistics.

Figure 7: Scattergraph to show results of the worked example



The main limitation of Spearman's Rank Correlation Coefficient is due to the ranking of the two data sets. This simply places the values in numerical order; it pays no regard to the magnitude of the differences between the values. An alternative statistical test of correlation, which uses the actual values rather than their rankings, so avoiding this pitfall, is the Pearson Product-Moment Correlation Coefficient. It is more sophisticated than the Spearman R test, but it does make an assumption that the data is normally distributed, which may not be the case. For Spearman, the data need not be distributed normally. This makes it suitable for a broader range of geographical investigations.

Focus Questions

1. Practise using Spearman's Rank Correlation Coefficient on data you may already have collected on fieldwork.

2. In what circumstances is Spearman's R not suitable in data processing? (This will make you think carefully about appropriate times to use this statistical method!)