Spearman's Rank Correlation Coefficient

The learning objectives of this unit are to:
* introduce Spearman's rank correlation coefficient as an alternative measure of correlation to the PMCC;
* to understand the relative merits of the two measures of correlation;
* to calculate Spearman's rank correlation coefficient.

Recap: The product-moment correlation coefficient is used to measure the strength of the *linear* association between two variables, i.e. how close the points on a scatter graph lie to a straight line. It is most appropriate when the points on a scatter graph have an elliptical pattern.

The product-moment correlation coefficient is less appropriate when the points on a scatter graph seem to follow a curve or when there are outliers (or anomalous values) on the graph:
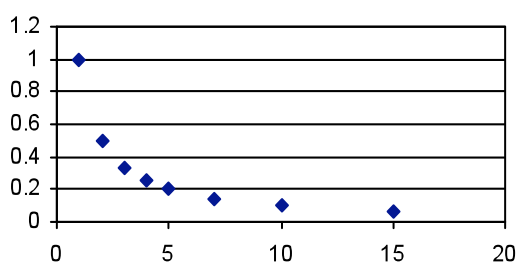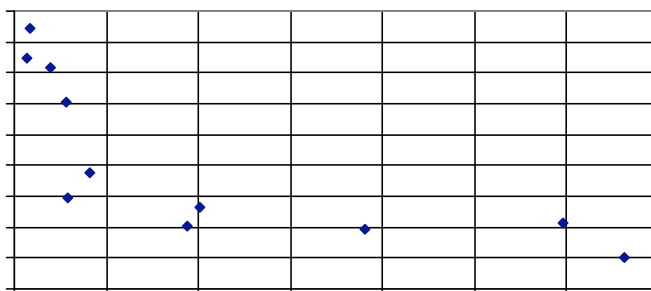

Introductory example:

The following data shows the annual income per head of population, *x*, (in US $) and the infant mortality, *y*, (per thousand live births) for a sample of 11 countries:

Country	A	B	C	D	E	F	G	H	I	J	K
x	130	5950	560	2010	1870	170	390	580	820	6620	3800
y	150	43	121	53	41	169	143	59	75	20	39

Scatter graph showing infant mortality

There is another type of correlation coefficient however called **SPEARMAN'S RANK CORRELATION COEFFICIENT** (denoted r$_S$).

One method of calculating Spearman's rank is demonstrated below:

Step 1: Rank both sets of data:

Country	A	B	C	D	E	F	G	H	I	J	K
x	130	5950	560	2010	1870	170	390	580	820	6620	3800
y	150	43	121	53	41	169	143	59	75	20	39
Rank *x*	1	10	4	8	7	2	3	5	6	11	9
Rank *y*	10	4	8	5	3	11	9	6	7	1	2

Note: In ranking the numbers above, I've used rank 1 to denote the smallest number in each row and rank 11 to represent the largest number. Some people would use rank 1 to denote the largest number and rank 11 to represent the smallest. It doesn't matter which way you do the ranking as long as you rank in the same way for both rows.

Step 2: Calculate the product moment correlation coefficient of the ranked data.

$$\sum x = 1 + 10 + ... + 9 = 66 \qquad\qquad \sum y = 10 + 4 + ... + 2 = 66$$

$$\sum x^2 = 1^2 + 10^2 + ... + 9^2 = 506 \qquad\qquad \sum y^2 = 10^2 + 4^2 + ... + 2^2 = 506$$

$$\sum xy = (1 \times 10) + (10 \times 4) + ... + (9 \times 2) = 293$$

So,

$$S_{xy} = 293 - \frac{66 \times 66}{11} = -103$$

$$S_{xx} = 506 - \frac{66^2}{11} = 110$$

$$S_{yy} = 506 - \frac{66^2}{11} = 110$$

Therefore,

$$\text{r}_S = \frac{-103}{\sqrt{110 \times 110}} = -0.936 \quad \text{(to 3 s.f.)}$$

Note: Remember to give answers to at least 3 significant figures.

Interpretation: This represents strong negative rank correlation between income and infant mortality, i.e. infant mortality tends to fall as income per head of population increases.

Note: There is a simpler way of calculating the Spearman coefficient when there are no tied ranks:

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

where *d* is rank *x* – rank *y* (i.e. the difference in the ranks) and *n* is the number of data pairs.

Country	Rank x	Rank y	Difference, d	d²
A	1	10	-9	81
B	10	4	6	36
C	4	8	-4	16
D	8	5	3	9
E	7	3	4	16
F	2	11	-9	81
G	3	9	-6	36
H	5	6	-1	1
I	6	7	-1	1
J	11	1	10	100
K	9	2	7	49
Total				**426**

$$\sum d^2$$

So,

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 426}{11(11^2-1)} = 1 - \frac{2556}{1320} = -0.936$$

Example 2: The table lists 7 schools and provides data about the percentage of pupils who have free school meals and their GCSE results. Calculate Spearman's rank correlation coefficient and explain what this means in the context of this question.

School	% of pupils claiming free school meals, x	% of pupils gaining 5 or more GCSEs at grades A*-C, y
Appledore	14.4	54
Butterscotch	7.2	64
Copperdale	27.5	44
Damson View	33.8	32
Eagle High	38.0	37
Forrest Green	15.9	68
Greengage	4.9	62

Solution: First rank the data:

School	Rank x	Rank y	d	d²
Appledore	3	4	-1	1
Butterscotch	2	6	-4	16
Copperdale	5	3	2	4
Damson View	6	1	5	25
Eagle High	7	2	5	25
Forrest Green	4	7	-3	9
Greengage	1	5	-4	16
Total				**96**

Using the formula

$$r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 96}{7(7^2-1)} = 1 - \frac{576}{336} = -0.714$$

This value represents negative rank correlation, so the schools with the highest proportion of pupils receiving free school meals tend to have the least successful GCSE results (and vice versa).

Example 3: Two students are considering applying to the same six universities (A, B, C, D, E, F) to study Zoology. Their orders of preference are as follows:

Student 1:   B   E   A   F   D   C
Student 2:   F   C   A   B   D   E

Calculate Spearman's Rand Correlation Coefficient and interpret your answer.

Solution: We start by presenting the data differently:

Student 1 put university B as 1st choice

University F was 4th choice

University	A	B	C	D	E	F	
Student 1, x	3	1	6	5	2	4	
Student 2, y	3	4	2	5	6	1	
d	0	-3	4	0	-4	3	
d^2	0	9	16	0	16	9	$\sum d^2 = 50$

So, $r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 50}{6 \times (36-1)} = 1 - \frac{300}{210}$

$$= -0.429$$

Interpretation: The students are in slight disagreement over their university preferences.

Note 1: If the orders of preferences were:

Student 1:   B   E   A   F   D   C
Student 2:   B   E   A   F   D   C

then $r_S = 1$ (total agreement between the students).
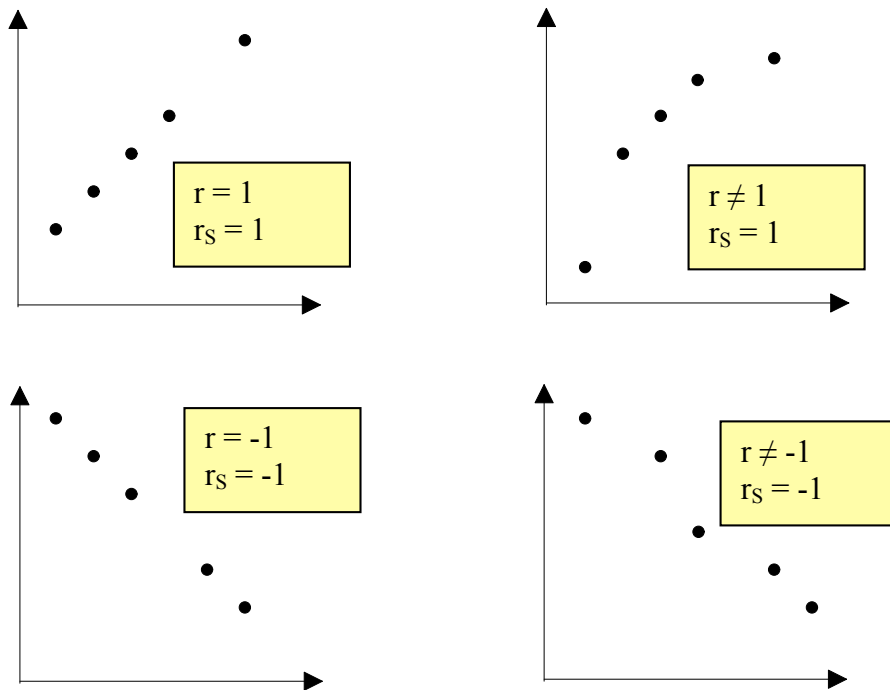
Note 2: If the orders of preference were:

Student 1:   B   E   A   F   D   C
Student 2:   C   D   F   A   E   B

then $r_S = -1$ (students hold completely opposing views).

r measures how close points on a scatter graph are to a straight line (i.e. the strength of a linear relationship).

r_S measures the tendency for y to increase (or decrease) as x increases, but not necessarily in a linear way.

Note: The PMCC is usually the preferred measure of correlation if the data has a linear relationship.

Effect of scaling on the PMCC

Consider two variables *x* and *y*. If new variables *u* and *v* are defined as:
     u = ax + b
     v = cx + d   where *a, b, c, d* are constants
then the correlation between *x* and *y* is the same as the correlation between *u* and *v*.

Example: Suppose that from the variables *x* and *y* we define two new variables:
     u = 2x + 5
     v = 0.5y - 7
The correlation between *x* and *y* is exactly the same as between *u* and *v* – in a scatter graph, the relative positions of the points is not changed by **linear scalings**.

Example 2: Suppose that from the variables *x* and *y* we define two new variables:
     $u = 2x^2 - 5x$
     v = 1/y
The correlation between *u* and *v* is **not** the same as between x and y as the transformations are **not linear**.