

Type of question: association between two dependent variables.

'Within'-subjects i.e. all subjects provide two measures.

Type of data: ordinal / non-parametric (Spearman rank correlation coefficient)

or interval/ratio / parametric (Pearson Product moment correlation)

Parametric tests can be used:

1. If data are on interval / ratio scale
2. There is a normal distribution in the data.

(bulk of distribution in the middle, no severe skew)

3. There is homogeneity of variance

If one of the DVs is parametric, but one is not,

you have to use the non-parametric test.

\* A correlation coefficient indicates the association between variables.

\* Correlations run from -1 to 1.

1 : perfect positive correlation

-1: perfect negative correlation

0 : perfectly no correlation

\* A correlation between 1 and about 0.6, or between -1 and -0.6 is strong.

\* Weaker correlations can still be significant (depending on the number of subjects), but may not be interesting.

Scattergrams: Y-axis: One DV, x-axis: other DV (handout)

## 1. Spearman's rank correlation coefficient

Example:

Is there a relationship between amount of sport taken and frequency of illness?

DVs: 1. rating for frequency of taking sport (1-10 never - every day)

2. rating for illness (1-5: very frequent - very rare)

Prediction: The more often sport is played, the rarer illness is.

H1: There will be a positive correlation between frequency of sport played and frequency of illness on the scales used.

H0: There will be no association.

12 subjects rate both measures:

Sub	Sport	ill
1	5	2
2	3	2
3	7	4
4	10	5
5	9	4
6	9	5
7	2	4
8	6	3
9	3	1
10	4	1
11	8	4
12	10	5

Does a high rank on one mean a high (low) rank on other?

Next stage: work out the rank that each score has, per dependent variable: (The scores can be on different scales)

\*Rank 1 is smallest

Example: in sport, 2 is the lowest score. It gets rank 1.

\*next score up is 3, but there are two of them.

They would have ranks 2 and 3, but instead, add ranks together and divide by number of scores with that rank.  $((2+3) / 2) = 2.5$ .

\*next is the score 4, it gets a rank of 4 (next rank up from 2 & 3, which were just allocated)

Do this for the whole set of scores, on both dependent variables.

Sub	sport	ill	rank-s	rank-i
1	5	2	5	3.5
2	<b>3</b>	2	<b>2.5</b>	3.5
3	7	4	7	7.5
4	10	5	11.5	11
5	9	4	9.5	7.5
6	9	5	9.5	11
7	<b>2</b>	4	<b>1</b>	<b>7.5</b>
8	6	3	6	5
9	<b>3</b>	1	<b>2.5</b>	<i>1.5</i>
10	4	1	4	1.5
11	8	4	8	7.5
12	10	5	11.5	11

At the next stage, calculate the difference between rank (sport) and rank (ill) for each pair of datapoints.

Sub	sport	ill	rank-s	rank-i	diff
1	5	2	5	3.5	1.5
2	<b>3</b>	2	<b>2.5</b>	3.5	-1
3	7	4	7	7.5	-0.5
4	10	5	11.5	11	0.5
5	9	4	9.5	7.5	2
6	9	5	9.5	11	1.5
7	<b>2</b>	4	<b>1</b>	7.5	-6.5
8	6	3	6	5	1
9	<b>3</b>	1	<b>2.5</b>	<i>1.5</i>	<i>1</i>
10	4	1	4	1.5	2.5
11	8	4	8	7.5	0.5
12	10	5	11.5	11	0.5

In this case, a low difference in rank supports the prediction, i.e. low rank sport (meaning sport is never played) is associated with a low rank in illness (meaning illness is frequent).

The next stage, is to square all the differences in rank

(to get rid of the minus signs)

Sub	sport	ill	rank-s	rank-i	diff	sq-diff
1	5	2	5	3.5	1.5	2.25
2	<b>3</b>	2	<b>2.5</b>	3.5	-1	1
3	7	4	7	7.5	-0.5	0.25
4	10	5	11.5	11	0.5	0.25
5	9	4	9.5	7.5	2	4

6	9	5	9.5	11	1.5	2.25
7	2	4	1	7.5	-6.5	42.25
8	6	3	6	5	1	1
9	3	1	2.5	1.5	1	1
10	4	1	4	1.5	2.5	6.25
11	8	4	8	7.5	0.5	0.25
12	10	5	11.5	11	0.5	0.25
sum						61

Then with the sum of the squared differences, work out the Spearman rank correlation coefficient (rho), using the formula:

rho =

$$1 - [(6 * \text{sum of squared differences}) / (N * (N - 1))]$$

so:  $6 * 61 = 366$

$$12 * (12 - 1) = 12 * 11 = 132$$

$$366 / 132 = 2.77$$

$$\text{rho} = 1 - 2.77 = -1.77$$

Note: rho is a descriptive statistic.

This is a strong positive correlation. To look up its significance, consult a table in e.g. Green & D'Oliveira, using the number of subjects to find the critical value of rho at various levels of significance.

One-tailed and two-tailed tests:

If your prediction is in a certain direction, then use a one-tailed test

If it's for a general association (or difference), then use a two-tailed test.

In our example we predicted a positive correlation: one-tailed.

Critical values for N=12

one-tailed test

<b>p</b>	<b>.05</b>	<b>.025</b>	<b>.01</b>	<b>.005</b>
rho (crit)	.506	.591	.712	.777

two-tailed test

<b>p</b>	<b>.10</b>	<b>.05</b>	<b>.02</b>	<b>.01</b>
rho (crit)	.506	.591	.712	.777

If the correlation is negative, then ignore the minus sign when looking up rho(crit) in the table.

In our case, .79 is greater than .777, which gives us a significance level of 0.005 on a one-tailed test.

Note:

If the correlation is perfect, the differences in rank will be 0.

$$1 - (0/N(N-Squared - 1)) = 1 - 0 = 1$$

Important: Interpretation:

Does taking a lot of sport make you healthy?

Does being rarely ill make you take more sport?

Do 'taking a lot of sport' and 'being rarely ill' correlate with something else, which causes both?

No causal relationship can be inferred.

Spurious correlations.

## 2. Pearson Product moment correlation

Only use when requirements for parametric tests are met.

So: Is the measurement scale appropriate on BOTH measures?

Is there a (roughly) normal distribution on both?

Is the variance roughly equal on both measures?

You might need to explore the data first to establish this.

Output of Pearson Product moment correlation:

r (between -1 and 1)

Example:

Is there a relationship between ability in maths and ability in linguistics?

dependent variables: 1. score on maths test

2. score on linguistics test

Prediction: Scores on a maths test will show an association with scores on linguistics test.

H1: There will be a correlation (positive or negative) between maths and linguistics test scores.

H0: There will be no association.

Test on two different parametric scales. There was a higher maximum score on the maths test. This is OK for correlations.

The calculation of Pearson's r

<b>sub</b>	<b>Maths</b>	<b>ling</b>	<b>m*l</b>	<b>m-sq</b>	<b>l-sq</b>
1	32	17	544	1024	289
2	54	13	702	2916	169
3	68	14	952	4624	196
4	93	10	930	8649	100
5	87	16	1392	7569	256

6	24	7	168	576	49
7	49	6	294	2401	36
8	35	18	630	1225	324
9	97	19	1843	9409	361
10	62	13	806	3844	169
11	44	9	396	1936	81
12	73	12	876	5329	144
<b>sum</b>	<b>718</b>	<b>154</b>	<b>9533</b>	<b>49502</b>	<b>2174</b>

The formula for r is as follows:

Divide  $[N * (\text{sum of all maths} * \text{ling}) - \{(\text{sum of all maths}) * (\text{sum of all ling})\}]$

by

the square root of  $(N * [(\text{sum of maths-squared}) - (\text{sum of maths})^2] * [N [(\text{sum of ling-squared}) - (\text{sum of ling})^2])$

Note the difference between sum of maths-squared (49502) and (sum of maths)squared (718-squared)

This formula gives you a Pearson Product moment correlation of 0.28

(Just as well we've got SPSS).

The df is  $N-2$ , in this case  $12-2=10$

$r = 0.28$  is a *weak* positive correlation. It might still be significant, though. (Note r is also a descriptive statistic)

r (crit) for  $df=10$ , two-tailed (no direction predicted)

<b>.10</b>	<b>.05</b>	<b>.02</b>
.4973	.5760	.6581

So, there is no significant association between maths and linguistics ability in this study.

Problems with correlations:

\* Range restrictions (if scores are very uniform, or from a small range, then correlation can be masked)

\* Heterogeneous subsamples: (if population measured on two DVs has populations with different behaviours, then correlation may be masked).

E.g maths/ling male/female. If there is a positive correlation between maths and ling scores in males,

and a negative correlation in females,

then overall, there may be no correlation, even though there is a (more complex) relation between the variables.

Regression:

Closely related to correlation

Two DVs.

Use one DV (predictor) to predict the variation on the other DV (criterion)

We will only deal with linear regression. Curvilinear regression not included. See Howell.

Remember scattergrams:

X = one variable, Y = other variable.

Regression line

If all points fall in perfect relationship Y to X, then line of best fit simply goes through all points on scattergram.

E.g.  $Y = X$  or  $Y = 2X$  or  $Y = 3 + 2X$

If there is no straight line through all points, then calculate 'line of best fit'.

in  $Y = 3 + 2X$

3 is the intercept (the value of Y when X=0) and 2 is the slope of the line.

more generally  $Y = a + bX$ . (Y is predictor, X is criterion)

The best fitting line is the line that leaves the least error of prediction (aka 'residual')

We use a formula that minimises the distance between all the points and the points on the line. However, as with SD, we square the difference, else the sum of differences would always be 0.

(Just for info, don't bother copying down)

$a = (\text{sum of all } Y) - b * (\text{sum of all } X) / N$

$b = (N * (\text{sum of all } X * Y) - [(\text{sum of all } X) * (\text{sum of all } Y)]) / (N (\text{sum of all squared } X) - (\text{sum all all } X)^2)$

N = number of subjects

This allows you to work out a and b.

You won't need to know this formula, just so that you've seen it once. The full maths is in Howell, Chapter 9. SPSS can do it for you, though.

All you need to know is that the slope and intercept give you a line that minimises the error of prediction.

Say you have found that  $Y = 3 + 2X$

You can then take each X point, and work out how well the formula fits the Y from each subject.

If you have an X of 4.5 and a Y of 10,

then you can obtain the 'residual' by slotting the X into the regression equation, which gives you an 'expected' score.

$Y = 3 + 2X$  ; X is 4.5 ( $Y = 3 + 2 * 4.5$ ); Y should be 12.

Y obtained = 10

Y predicted = 12

10 is 2 away from 12, so for this pair of data point you have an error of prediction of 2.

If you square the correlation coefficient, you know what proportion of the variation on Y you can predict by knowing X.

A correlation of 1 or -1 allows you to predict 100% if the variation of Y given X.

Correlation of 0.8 or -0.8: .64

Correlation of 0.6 or -0.6: .36

So, even with a strong correlation, which may be significant, you cannot always predict a large proportion of the variation on Y given X.

Summary:

- \* Spearman (non parametric) and Pearson (parametric) correlations.
- \* Regression: predicting X from Y.
- \*  $r^2$  : How much of the variation on Y can be predicted from X.

KEY TERMS:

SCATTERGRAM, SPEARMAN RANK CORRELATION COEFFICIENT, PEARSON PRODUCT MOMENT CORRELATION, ONE-TAILED TEST, TWO-TAILED TEST, REGRESSION, LINE OF BEST FIT, PREDICTOR, CRITERION, SLOPE, INTERCEPT,  $r^2$